Foundations and Trends® in
Information Retrieval
Vol. 6, Nos. 4–5 (2012) 257–405

DOI: 10.1561/1500000023

now
the essence of knowledge

Contextual Search: A Computational Framework

By Massimo Melucci

Contents

Foundations and Trends® in
Information Retrieval
Vol. 6, Nos. 4–5 (2012) 257–405

DOI: 10.1561/1500000023

now
the essence of knowledge

Contextual Search: A Computational Framework

Massimo Melucci

Università di Padova, Dipartimento di Ingegneria dell'Informazione, Via G. Gradenigo, 6, Padova, 35131, Italy, massimo.melucci@unipd.it

Abstract

The growing availability of data in electronic form, the expansion of the World Wide Web (WWW) and the accessibility of computational methods for large-scale data processing have allowed researchers in Information Retrieval (IR) to design systems which can effectively and efficiently constrain search within the boundaries given by context, thus transforming classical search into contextual search. Because of the constraints imposed by context, contextual search better focuses on the user's relevance and improves retrieval performance, since the out-of-context aspects of the search carried out by users that are likely linked to irrelevant documents are left apart.

This survey introduces contextual search within a computational framework based on contextual variables, contextual factors and statistical models. The framework adopted in this survey considers the data observable from the real world entities participating in contextual search and classifies them as what we call contextual variables. The contextual variables considered are content, geotemporal, interaction, and social variables. Moreover, we distinguish between

contextual variables and contextual factor: the former is what can be observed, the latter is what cannot be observed, yet this is the factor affecting the user's relevance assessment. Therefore, in this survey, we describe how statistical models can process contextual variables to infer the contextual factors underlying the current search context.

In this survey we provide a background to the subject by: placing it among other surveys on relevance, interaction, context, and behavior; providing the description of the contextual variables used for implementing the statistical models which represent and predict relevance and contextual factors; citing and surveying useful publications to the reader for further examination; providing an overview of the evaluation methodologies and findings relevant to this subject; and briefly describing some implementations of contextual search tools.

1

Introduction

> Context: from Latin *contextus*, where *con* stands for "together" and *texere* stands for "to weave".
>
> Oxford Dictionary

1.1 Motivation of this Survey

Many researchers with various backgrounds believe that context can enhance the user's experience and improve the system's effectiveness of search. In so doing, they frame Information Retrieval (IR) within the more general notion of contextual search, although from differing viewpoints. The different perspectives at which context has been viewed have led to definitions of context with different potential of implementation.

At one extreme, context can be defined as the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed. From this perspective, some publications relevant to contextual search are being written mostly from an information seeking and retrieval point of view. Although such a point of view is rooted in strategically important disciplines like user behavior, cognition or human interaction, it cannot fully help see how to

proceed with what should be observed and computed for implementing context within an IR system. At the other extreme, context can be viewed as the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning.

These two extreme perspectives clearly differ in their potential of implementation. While the definition of context in terms of events and settings cannot obviously be utilized for designing algorithms and data structures, the definition of context as a text window around a word is easier to implement and is strictly related to the nature of text and is part of common sense. However, this view is quite reductive and considers only one of the ways context occurs.

From one extreme to the other, computational approaches to contextual search followed one another, ranging from sophisticated and computationally expensive approaches to more simple and efficient ones, but each one of them has been useful for writing this survey.

1.2 Definitions and Scope of the Survey

A *variable* is any observable value that is liable to change. Variables can be: qualitative or quantitative; ordinal or not; if ordinal, cardinal or not; if cardinal, integer or not; and so on. When the variables are random, they change according to a probability distribution in such a way that its observation value occurs with a given probability. Such a characterization allows inference to be made on estimation and prediction of potential relationships between variables in such a way that the variation of some independent variables determines the variation of some dependent variables.

A *contextual factor* is any unobservable circumstance or fact of search such as query intent, personal interest and document quality, which affects relevance. We concentrate on three contextual factors: query intent, personal interest, and document quality. (The contextual factors are illustrated in Sections 2, 3, and 4.)

Query intent refers to the objectives of the user who issued the query. In this situation, a query is viewed as a means to accomplish a task such as "dissertation writing," "finding a resource," "bibliography compilation" and a query intent is an objective to be achieved in order

to accomplish the task. Intent is a property of a query and is not necessarily tied to a user in the way a personal interest is.

Personal interest in general refers to the user's state of wanting to know or learn about a thing, a person or an event. An interest is an information need that has the quality of sparking curiosity or holding the user's attention, and may be viewed as a property of an information need that makes the information need crucial to the user.

Document quality refers to the property of a document that is able to be trusted as being up-to-date, authoritative, exhaustive, accurate, reliable, and clear. A high-quality document is considered to be the best of its kind and unlikely to be improved upon.

A *contextual variable* is any set of variables dependent on contextual factors. This survey classifies the contextual variables observed by the real world entities participating in contextual search as: content variables, interaction variables, geographical variables, and social variables. These contextual variables are introduced in Section 1.3.

A *statistical model* is a a set of computable mathematical rules defined over a set of variables or factors for example height and weight are related in a way that they can be plotted as points along a straight line, or the frequency of a term within a document and relevance assessment are related in a way that the higher the frequency, the more likely the document is relevant. In this survey, the rules of a statistical model express computable relationships between contextual variables and contextual factors.

The denotation of context in this survey is thus essentially computational and allows us to introduce a computational framework of contextual search summarized by the following

---

Definition 1.1. *Context* is represented as a set of contextual variables and contextual factors weaved together by statistical models of estimation and prediction.

An objective of this survey is to inform the reader which contextual variables, contextual factors, and statistical models have been utilized in the literature to represent context by means of a

computer system and therefore which of these provide some useful hints about what to use to extend a traditional IR system toward context.

The definitions of contextual variable and contextual factor point out two main aspects: observability and dependence. Observability is a necessary condition of the meaning of context in this survey, that is, context can be operationalized as variables and can therefore be exploited in search only if some variables can be defined and observed; context implementations not referred to as variables are not considered in this survey. A basic example is standard IR: index term occurrence is a contextual variable observed from the document content, in contrast, aboutness is a contextual factor affecting relevance and cannot directly be observed.

Dependence is between contextual factors, relevance, and variables. Research in IR often assumes that a variation of contextual factors reflects upon a variation of the contextual variables. The relationship posited between personal interest and term frequency is an example of the relationship between a contextual variable and a contextual factor; term frequency may increase in a document if this document becomes interesting for a user. Therefore, if some variations of the contextual variables are observed, a variation of the contextual factor is likely to have occurred; for example, if term frequency is higher in a document than in another document, the former is more likely interesting than the latter.

In the computational framework presented in this survey, attention is also paid to discovering the contextual factors that affect relevance assessments. Thus, a variation of relevance assessment is due to a variation of the contextual factors; for example, if query intent is viewed as contextual factor and term frequency is a contextual variable, a variation of frequency may result from a variation of query intent which in turn affects relevance.

Another feature of this survey is the attention paid to statistical models. The statistical models mentioned in this survey provide some advantages in illustrating context. They allow researchers to implement context because these models are suitable for estimating and predicting context starting from the variables observed in objects. Another

advantage is of a computational nature. Most of the statistical models scale up when the size of data available from the user's environment, social network, and personal dimension increases by hundred-fold, thus keeping high levels of effectiveness and efficiency. Everyone prefers computationally efficient approaches to search in context; however, computational efficiency is not a feature of every one of the approaches (e.g., those arising from Artificial Intelligence), despite the computational potential that makes contextual variable implementation more unbridged than in the past.

A computational framework for contextual search like that described in this survey may resemble classical modeling in noncontextual search in which the "best" model is selected for optimizing effectiveness. We think that this approach is not constrictive since in the past there have definitely been statistical models for contextual search that resulted in significant improvements in IR systems. We do not claim that a computational approach is the only approach to explaining contextual search, but we do claim that it is the best approach for making contextual variables useable; relevance feedback is an example of a computational approach to contextual search thoroughly investigated in the past.

Although (or maybe precisely because) investigated and employed for a long time, relevance feedback is still crucial in contextual search, since it mainly relies on content variables and in particular on document content. Indeed, relevance feedback has recently been reevaluated and experimented with huge test collections and very short noisy queries through initiatives such as the relevance feedback track of TREC. However, despite it being relevant, explicit or pseudo-relevance feedback is not addressed in this survey because our focus is on recent developments of contextual search while there are already surveys of relevance feedback and query expansion.

In contrast, implicit relevance feedback is the backbone of the incorporation of behavior in contextual search. The research conducted within implicit relevance feedback has aimed to use the contextual information generated during the interaction between the user and information as implicit evidence of relevance. Hence, a key question is whether implicit relevance feedback can effectively be used in

contextual search systems in comparison with traditional content-based ranking functions or more advanced yet well experimented methods such as anchor text or link analysis algorithms.

As it happens, contextual search is not relevant only to IR, of course, but to other research areas too with which they interact yet they seem rather distant from IR. Examples are Psychology, Mobile Communication, Electronic Commerce, Nomadic Computing, Human Computer Interaction. All these subjects are relevant to this survey although they cannot be looked at thoroughly because the topic of this survey is already vast enough. A few things that are on the side of the context of a document and are not the primary focus are: temporal context (e.g., two e-mail messages sent right after the same event); storage context (e.g., two documents found in the same file system folder); conversational context (e.g., one e-mail message is a reply to another).

1.3 Contextual Variables

This survey considers four types of contextual variables: content, geographical, interaction, and social variables.

Content variables refer to the informative content and relationships of queries and documents. The data are content features observed from text, image, video, audio; link anchors; layout; genre; lexical properties (e.g., part-of-speech tags); user's tags (e.g., image tags or file names); category labels (e.g., Wikipedia category labels); demographic labels (e.g., authorship) and anything used to describe informative contents or to enrich information need representations.

Geographical variables are any variable with the state of existing within or having some relationship with space location. Examples are geographical names added to documents or queries, digital photographs tagged with geographical coordinates, typically the latitude and longitude of the space location perhaps associated to a user.

Interaction variables are observed over time during the interaction between users and IR systems. (Geographical variables are not necessarily referred to a user.) These variables are for example: click-through data; data about queries or search sessions; user

judgments or assessments; user behavior data (e.g., document retention, display time, eye or mouse movements).

Social variables refer to user communities or groups and are observed for example from: "tweets"; social connections (e.g., friendship); hyperlinks (e.g., a link between two WWW pages).

1.3.1 Content Variables

Content is a contextual variable exploited in contextual search to decide whether an additional or special action that is different from time to time should be performed by an IR system when the user is interacting with the informative content managed by the system for meeting his information needs. Content variables can be observed from the documents of a collection, search engine result pages, queries, or from parts of them such as windows, fragments, and passages.

The main *medium* addressed in the literature of this survey is text. It is perhaps the richest source of evidence for predicting context since text is an expression of natural language, that is, the main means used by humans to communicate information and needs. Text can easily be managed because words or terms can be suggested to the user who in turn can understand them by leveraging common cognitive abilities and feed data back into the system: positive words represent what items the user would like to retrieve; negative words indicate what the user does not want; neutral words are not good indicators of her information needs. We are not dealing with multi-lingual text IR; the literature utilized in this survey refers to the English language only.

1.3.2 Geographical Variables

In our view, geographical variables are observed and are instrumental for detecting contextual factors such as query intent and personal interests. Geographical variables differ from other variables due to the intents underlying the queries referring to geographical information. However, they raise issues similar to the issues raised by natural language processing. In particular, when geographical variables are names, the issues are: name or reference detection,

name disambiguation, name clustering, linking or association, name weighting, and document ranking.

Our use of geographical variables complements the view of geography as a relevance factor; for example, Raper [143] defines geographical information needs based on cognitive and geographic criteria and argues that geographic relevance is best defined as a spatio-temporally extended relation between geographic information needs and geographic documents.

1.3.3 Interaction Variables

Whenever users have difficulty in expressing their information needs, contextual variables based on interaction are precious because an IR system can be enabled to automatically deduce a user's interest based on the data generated during the interactions with the system. Indeed, the data observed over time during the interaction of the user with a contextual search system form an interaction history where history also means "finding out." Thus, the value of interaction variables is not only the individual pieces of content, but their organization within a coherent stream of data — it is the observation of these pieces together which makes history valuable; for example, if the user has requested some documents recently, it is likely that the user is in a given context, and the retrieved documents can form the basis of supervised learning for the user's preferences because of recency and not only because of the amount of data.

When the interaction data employed for estimation are very close in time to the user's actions, the estimated models are more closely related to the user than the models that would be estimated with the farthest data in time. The data employed for estimation are very close in time to the user's actions when, for instance, it is of interest to the contextual search system to recognize the correct query sense or intent of the user. On the other hand, implicit relevance feedback information collected over a long period of time is less likely to be very useful for predicting the individual's interests than the immediate search context and feedback information; they may be useful for predicting the interests of a group.

1.3.4 Social Variables

What a user either directly or indirectly learns from or teaches to the communities is a crucial contextual variable because human relationships are at the basis of many conditions. Hence, it is not surprising that quite an appreciable proportion of the literature on contextual search addresses the issues of the social dimensions of the users. Community is meant in a broad sense and is not confined to social networks or similar user organizations; the same algorithms can make the social variables observed from any user community a useable source of evidence for contextual search.

In this survey, we consider some cases of community behaviors where members participate in a collective activity and unwittingly collaborate to build collective knowledge. These kinds of community behaviors differ in the degree to which a member is aware of belonging to a community. For example, the users tagging resources are often aware of their membership to a community (e.g., they log into a system) whereas the users clicking ads are not aware that their clicks are collected and exploited for boosting ad ranking. What the various kinds of community behaviors have in common is that they leverage the (large) size of the communities involved in a way that the large quantity of observed data can be exploited to estimate parameters and discover patterns useful for implementing contextual search.

In IR and related disciplines there have already been research works that to a certain extent investigate how members of a community interact, perhaps indirectly, to building knowledge that is further exploited by the community (the link analysis methods addressed in this section are an example and the earlier bibliometrics is another notable example).

Numerous papers addressing social variables as contextual variables are based on link analysis algorithms since a graph is a natural way to represent a community; nodes are members and edges are relationships between members. However, in this survey, we not only address link analysis but also address other statistical models suited to mining useful information from community contextual variables. To this end, we are drawing the reader's attention to a couple of social variables,

that is, "tweets" and tags, which can efficiently provide useful information about the context on a large scale.

Another research area is known as digital annotation systems. Annotations affixed to digital documents is a little more recent than bibliometrics because the use and production of the digital documents has grown since the 1980s at the earliest; Agosti et al. [5] introduced digital annotation systems. However, since their advent their use is still limited, thus making the exploitation of these data for contextual search through statistical methods difficult. We focus on two types of annotation (ESP games and "tweets") that in contrast to "traditional" annotation affixed to digital documents stimulate the implementation of large scale statistical methods for contextual search.

1.4 Historical Background

In this section we provide a background to contextual search by: placing the subject among other surveys on relevance, interaction, context, and behavior; citing and surveying useful publications to the reader for further examination.

Before the relatively renowned and growing interest in contextual search viewed in the recent literature of IR, context had been on the scene, or perhaps better stated behind the scenes, for many years (perhaps for decades) as the IR literature since the 1970s shows. As the literature is by now quite vast, we can distill only some aspects and issues and cannot be more exhaustive than the publications already available on this topic.

This section is then devoted to providing a summary of and the references to the publications in which contextual search has been thoroughly considered. These publications may provide the reader with complementary information, and give a background to this survey. In particular, this section draws the reader's attention to the papers by Belkin et al. [18]; Ingwersen and Järvelin [78]; Mizzaro [131]; Ruthven [149]; Saracevic [152]; Spink [159].

1.4.1 Relevance

Because users of an IR system assess whether a document is relevant in a context, context has been a crucial aspect of relevance for decades.

Hence, relevance is intrinsically dependent on context. Due to the complexity of context and relevance, the most common IR models are a mere simplification of the reality in which users are called on to assess the relevance of documents to their information needs.

If the items of a context are gathered together, a sort of relation is obtained; actually, a mathematical relation as it is intended by a DBMS. Saracevic [153, p. 1918] suggested an understanding of relevance as a relation. According to this understanding, relevance is a relation over information objects and contexts which include information needs, tasks, and other elements. In Saracevic's review, context is an element of relevance ("Relevance has a context") and it is viewed as a complex, dynamic "interaction between a number of external and internal aspects, from a physical situation to cognitive and affective states, to motivations and beliefs, to situations, and back to feedback and resolution." Context is "ambiguous, even amorphous" and at most "context is a plural."

In the review of relevance authored by Mizzaro [131], context "includes everything not pertaining to topic and task, but however affecting the way the search takes place and the evaluation of results." This definition suggests the view that the user has some context that is not stated in the query but which we could nonetheless model. Mizzaro's paper also cites literature relevant to context introduced as a factor, component or container of the content, user, task, and so on.

1.4.2 Anomalous State of Knowledge

The Anomalous State of Knowledge (ASK) by Belkin et al. is another useful element for understanding contextual search. The first part of the paper reported by Belkin et al. [18] introduces the ASK hypothesis stating "that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly" [18, p. 62]; the second part reported by Belkin et al. [19] describes an experiment. The information need of the ASK hypothesis stems from a "topic or situation" which might better be named as problematic situation or task. In the words of Belkin et al., "the user, faced with a problem, recognizes that her/his state of

knowledge is inadequate for resolving that problem, and decides that obtaining information about the problem area and its circumstances is an appropriate means toward its resolution."

When the ASK hypothesis is valid, the user is unable to make his information need explicit because what he would be asked to say is precisely what he does not know. A consequence of this impossibility which is relevant to this survey is that, to address the ASK, an IR system should be interactive and iterative, thus calling into play various contextual factors such as query intents, personal interests, and document qualities. Sometimes, the combination of different contextual variables leads to concept networks. Belkin et al. [18, p. 68] defined concept networks as networks of inter-related documents and named them as "formal context." Such a network becomes a description of context and at the same time a source of evidence from which data can be observed to represent context. Networks of concepts have been further elaborated in Agosti et al.[4] within the most naturally interactive system, that is, hypermedia systems.

1.4.3 Interactive Information Retrieval

Ingwersen and Järvelin [78] introduced the Integrated Cognitive Research Framework for IR. The components of this framework are: information objects (e.g., documents); the IT component (e.g., search engines); the interface (e.g., WWW clients); the cognitive actor (e.g., the user); the socio-cultural and organizational context (e.g., the workplace or the community). Between the components, which are depicted in Figure 1.1, there are influence or exchange relationships depicted as unidirectional and bidirectional arrows, respectively, and there are solid or dashed unidirectional arrows corresponding to influence and influence over time, respectively. Within the Integrated Cognitive Research Framework for IR, the definition of context suggested in Ingwersen and Järvelin [78] becomes: "in information seeking and retrieval actors and objects [are] associated with each component of the cognitive information seeking and retrieval framework function as context for their own elementary cognitive structures (intra-object context), as context to one another (inter-object context), and in context of

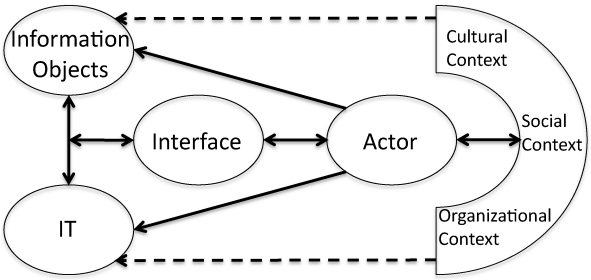

Fig. 1.1 Ingwersen and Järvelin [78]'s Integrated Cognitive Research Framework for IR.

the interaction processes between framework components, which themselves are contextual to each other. In the latter case one may talk about social/organization/cultural as well as systemic contexts. The context of interactive IR processes ranges from algorithmic IR processes in context of interactive IR as well as information seeking processes to information behavior. All information seeking and retrieval components and activities are in context of common social, physical and technological infra-structures as well as their history over time."

As it happens, circumscribing a notion like that of context to something simpler and perhaps simplicistic makes its implementation easier or more understandable than general or perhaps vague definitions. An example is a user interface-oriented notion of context which would help visualize the different components of Ingwersen and Järvelin's framework. Ruthven [148] gives a user interface-oriented notion and states that "Our ideas on context (from both a soft and hard laboratory perspective) often manifest themselves at the interface." Lalmas [106] adopted this definition.

Ruthven [149] some years later provides another definition: context is "a complex set of variables describing our intentions, our personal characteristics, the data and systems available for searching, and our physical, social and organizational environments" or it is also thought as the fact that "personal [context] information can cover any information that we have experienced (such as webpages we have visited), information that we have received (such as email) or information that we have created (such as documents or images). [...] [T]he range of contextual factors that might be important is vast ranging from age, physical

and cognitive ability (which may require altering the presentation of search results as well as the selection of results), learning styles, education level, or mood of searcher. The most common personal context investigated so far is the searchers topical search interests, particularly through applications of information filtering."

To personal context, we may add social context, which is somehow related yet independent of other contexts since it is about "how people use systems and for what purposes. We can mine this information — the context of use — for many purposes including filtering information to obtain better search results" according to Ruthven [149]. From this point of view task is the information problem, for example, finding a holiday destination, writing an essay, giving a lecture, which is the reason why the user expresses his information need through queries, browsing, clicking, etc. Thus, task context covers any information that describes the user's problem and that makes relevance, usefulness or authoritativeness of documents dependent on the task, with all the other variables being equal.

Space–time reality is perhaps the most intuitive and common setting where we experience context. Thus, it is quite straightforward to define physical context as the container of important data for providing situationally relevant information (e.g., GPS coordinates or time). Similarly, environmental context relates to any information about the type of location where the user's search takes place (e.g., whether the user is in a public place, the weather is nice, the roads are congested) according to Ruthven [149].

Contextual search can barely be separated not only from IR and information seeking and retrieval but also from the notion of human information behavior defined by Spink [159] as follows: human information behavior "refers to a wide range of processes which people employ when engaged with information and to related cognitive and social states and effects". In a sense, human information behavior studies are orthogonal to ASK, IR and information seeking and retrieval since they aim at understanding how and why the users interact with information when this information is contained in documents or queries. Spink in particular is interested in the user's behavior during the formulation of the ASK. She defines information seeking and retrieval

as "one sub-process within human information behavior that includes the purposive seeking of information in relation to" an ASK because information seeking and retrieval starts when an ASK has been recognized, continues when relevance assessments have been observed and ends when the ASK has been solved. From this description it is then clear that information seeking and retrieval is as highly dependent on context as human information behavior is. The remarks made by Spink [159] about human information behavior within communities and the personal dimension are relevant to this survey.

1.5 Concluding Remarks and Suggestions

The computational framework underlying contextual factors, contextual variables, and statistical models is the main conceptual contribution of this survey. Other researchers are allowed to place other contextual factors, contextual variables and statistical models in this framework, thus preserving the overall consistency of the illustration of contextual search proposed in this survey. Some results illustrated in the remaining sections may well be placed in more than one contextual factor (e.g., understanding the intent of a given user may be placed in Section 2, in Section 3, or both). However, these decisions are a matter for the researchers implementing this framework. Appendix A briefly illustrates some prototypes of contextual search tools.

We conclude by giving some bibliographic references relevant to the computational framework introduced in this survey and to the general notion of contextual search. Alpaydin [7] describes support vector machines. Alpaydin [7] is a reference on machine learning. Azzopardi [12] gives a thorough study that starts from theoretical issues, investigates whether and how language models can be an efficient and effective theoretical framework for contextual search, and ends with experiments. Bai et al. [14, 15] are examples of text window-based context papers with co-occurrence analysis, an interesting modeling of contextual factors based on language models and an analysis of domain knowledge and language model combination. Bartholomew et al. [17] provide a perspective of the factorial models that are relevant to the notion of computational

framework used in this survey. Bian et al. [23] are worth reading as for the Expectation-Maximization algorithm. Blei et al. [24]'s is the original publication on latent Dirichlet allocations. The notion of geographical variable is discussed by, for example, Cai [33]. The remarks made by Chakrabarti et al. [41] on how to build an effective model and avoid bias, overfitting, etc. are useful to a newcomer to machine learning because they explain basic issues in a realistic scenario. Croft and Lafferty [47] survey language models for IR. The study by Efthimiadis [57] describes query expansion whereas the paper by Carpineto and Romano [38] is an up-to-date survey of this topic. Feller [60]; Levinson et al. [113]; Rabiner and Juang [140] are some reference publications on Bayes' rule, Markov chains and hidden Markov models. Halmos [68] explains Singular Value Decomposition and in general vector spaces. The paper written by Hu et al. [75] is easy to read and has a computational flavor. As for interaction variables, the reader may want to spend some time reading Inmon [79, 80] who introduced the notion of time-variancy, since click-through datasets may be viewed as an instance of data warehouses. The special journal publication edited by Jones and Purves [90] is a useful reference on the issues of geographical variables. The papers on implicit relevance feedback by Kelly and Belkin [95, 96]; Kelly and Fu [97]; Kelly et al. [98]; Kelly [92, 93, 94] are definitely worth reading. The survey by Lalmas and Ruthven [107] provides a precise, recent and exhaustive account of relevance feedback. Lau et al. [108] address context at difference abstraction levels, from the conceptual, to the logical up to the statistical level. Lau et al. [109] present an interesting application of their theoretical framework and show that the vector space model is still a good baseline for search in context. Metzler and Croft [130] illustrate conditional random fields. Ponte and Croft [138] introduce language models for IR. The notion of geographical variable is also discussed by Reichenbacher [144]; Reichenbacher and De Sabbata [145]. The paper written by Shannon [155] is the reference for entropy. The papers on exploratory search by White et al. [172]; White and Kelly [173]; White et al. [170, 171]; White and Roth [174]; White [169] are also useful reading.

2

Query Intent

Query: from Latin *quaerere*, "ask, seek."
Intent: from Latin *intendere* ("intend," "direct"), from
in- "toward" + *tendere* "stretch," "tend."
Oxford Dictionary

2.1 Introduction

A user starts searching because he is often in charge of accomplishing a task. The significance of task in IR derives from its role in explaining the differences in relevance assessments and then in designing systems. Task has been recognized as one of the main contextual factors that give rise to the user's ASK (Anomalous State of Knowledge) according to Belkin et al. [18, 19]. Therefore, detecting task helps retrieve relevant documents. From an information seeking and retrieval point of view, the recent literature devoted to the influence of task on contextual search mainly reports naturalistic user studies. An example of these studies is reported by Kelly and Belkin [96] where the authors report an understanding of the subjects' natural searching behaviors.

In this section, we are distinguishing between search task and query intent. While accomplishing a task, the user decides intents, and he may

change his intent while executing the task. Thus, query intent is what the user directs the attention to, while search task (or task, in brief) is a search to be done or undertaken.

In parallel to the approaches to task detection proposed from an information seeking and retrieval point of view, the issue of query intent detection has often been addressed from an IR point of view as a problem of query disambiguation of which solutions have frequently been based on query expansion and relevance feedback.

Due to its fame and effectiveness within many situations, it is not surprising that query expansion has been investigated and has enjoyed a revival in designing contextual search systems. When query expansion is exploited in contextual search, the sources of evidence are mainly provided by documents, queries, and relevance assessments, whereas the main statistical models belong to a vast class of relevance feedback methods. The methods for query expansion have been investigated and employed in IR for some decades in various forms (i.e., pseudo- or implicit feedback) depending on the particular retrieval problem or constraints.

In this section, more recent and advanced methods for detecting query intents are addressed in addition to the most common, query expansion-based methods. Often, these advanced methods rely on some sort of query intent taxonomy although the definition of a query intent taxonomy would rather be difficult and would yield to an elaborate taxonomy attempting to capture many detail of intents.

2.2 Is Query Intent Predictable?

The basic question preceding how one can identify the query intent automatically without any explicit feedback from the user is whether queries have a predictable query intent; here, predictable means that it is possible to have a method that associates a query with a particular intent by only looking at the query features.

Predictability mainly depends on whether an intent does actually exist. In fact, the intent of a query is not always predictable since it does not exists as frequently as one may expect. Although an intent almost always implies a query as already explained by Belkin et al. [18, 19],

strangely enough, an intent can be predicted only half of the times. Lee et al. [110] show that intents can be predicted for 60% of the queries. The majority of these queries are informational while the intent of the other 40% of the queries, which are likely navigational, may be detected with *ad-hoc* methods.

Query intent is often neither navigational nor informational, but simply to get access to an online resource as reported by Rose and Levinson [147]. They indeed distinguish between navigational queries, which aim at locating a WWW site, from resource finding queries, which aim at obtaining a specific resource (e.g., digital file, e-mail address, etc.), the latter distinction having been made also within the WWW track of TREC that planned two distinct tasks, that is, homepage finding and resource finding since they may require distinct retrieval techniques.

Many informational queries that attempt to locate a product or service rather than to learn about it should be added to the 40% of navigational or resource finding queries as reported by Rose and Levinson. It is quite surprising, indeed, that about one third of all queries appeared to be informational for which traditional IR systems were designed. These percentages have to be compared with those reported by Broder [26] where it is stated that the percentage of navigational or transactional queries is between 40% and 50%.

According to Gan et al. [63], 13% of the queries have a geographical intent, that is, the user's intent is navigational or informational yet refers to close locations. Moreover, the probability that a query has a geographical intent when it contains a geographical name is only 33%. The good news is that only 1% of queries has a geographical intent when they do not contain any geographical name.

2.3 Detecting Query Intents Using Interaction Variables

2.3.1 Using Click-Through Data

Query intent may be learned from how users have interacted in the past with the returned results for this query. Computing statistical distributions of click-through data was studied by Lee et al. [110] who were among the early researchers who postulated a relationship between

query intent and click-through data. The basic rule, which has also been tested in other papers, is that navigational queries are followed by one click while an informational query is followed by a series of clicks. In particular, Lee et al. suggested applying statistical moments of a distribution and investigating anchor-link distribution other than the more commonly investigated click-through data distributions. Lee et al. associate a normalized number of distinct WWW pages to each query that occurs as anchor too. Low values of this number signal a navigational query because a few yet frequent pages are visited after the query is issued.

A relationship was found between query intent and query frequency. In particular, Downey et al. [55] found that users behave according to query frequency or to URL frequency — for example, search session length increases when the query is rare. These results may be due to the underlying query intent which influence both query frequency and user behavior.

There is not only a relationship between query intent and search session length; query intent evolves over time, as reported by Kulkarni et al. [105]; in particular, they found that popularity features possibly indicate a change of query intent. Moreover, query popularity and intent are correlated, since popularity increase signals the query is becoming informational. The findings reported by Kulkarni et al. [105] include a taxonomy of query intent change; we think this taxonomy should be paired with the taxonomies of query intents proposed by Broder [26]; Broder et al. [27]. In particular, Kulkarni et al. suggested that changes in query intent are due to the user's need of zooming (i.e., making the intent more or less specific) or shifting (i.e., moving the intent toward another target).

However, some problems may arise when using click-through data for detecting intents. Only relatively few queries are issued multiple times, while there are many rare queries without sufficient click-through data. The fact that click-through data are not always good predictors especially when queries are rare has been confirmed within contextual advertising by Ashkan et al. [10]. They found that the click-through rate is highest for commercial queries that are also navigational queries. However, this finding should be cleansed of

the effects of the placement of advertisements, since this placement influences on the number of clicks the advertisement receive. Commercial queries are indeed very frequent. After cleansing the effects of the placement of advertisements, it was found that commercial-navigational queries have advertisements placed at different ranks of the search engine result pages with higher click-through rate than the advertisements associated with the commercial queries that are also informational.

The sparsity of click-through data is often the main problem underlying the research work aimed at detecting query intents. The click-through data used to detect contextual factors can well be different from classical query or document content, since click-through data are not intrinsically carrier of meaning as words are.

Click-through data may rely on more "algorithmic" sources of evidence, such as graph patterns. Graphs may provide richer information representations than "flat" statistical distributions. Computing patterns from graphs linking queries and clicked documents was described, for instance, by Li et al. [117] who start from the assumption that queries with similar click patterns are likely to have the same intent, therefore, detecting a query pattern is useful for detecting a query intent. A click pattern can be observed from the bipartite-graphs where the edges are between queries and WWW pages and can be weighted by click frequencies.

It is possible to manually label a small set of seed queries in the graph, and then propagate the labels to other queries as suggested by Li et al. The problems that have to be addressed include defining a stop criterion and a regulation mechanism. Regulation is necessary to limit the propagation of wrong labels. The algorithm that is illustrated by Li et al. recalls PageRank which has been used in IR. A bipartite click graph is described as a matrix such that an entry is the frequency of clicks from a query to a WWW page — there are as many rows as queries and as many columns as visited pages. The set of manually labeled seed queries are arranged in a matrix Q such that an entry is $+1$ (or -1) if the query that corresponds to the row is a positive (or negative) example for the label corresponding to the column of Q. Q thus includes the likelihood that a query belongs to a class and

it is iteratively updated by the algorithm to eventually output a final likelihood matrix (note that likelihood is different from probability). At each step, the algorithm computes another matrix L as the product of the two previously described matrices. An entry of L is the likelihood that a label refers to positive examples. L is finally mixed with the current likelihood and bipartite click-graph matrices where the mixing parameter plays the role of regularizer in the same way the damping factor does in the PageRank algorithm.

Label propagation was tested in two classification tasks reported by Li et al. Product intent refers to queries looking for any products or class of products which can be purchased in store or online. Job intent refers to queries for finding jobs. A dataset was prepared from a click-through data set available from a commercial search engine. From this click-through data set, seed query sets for manual labeling were selected and click graphs were constructed. Another set was prepared for performance testing.

The above described link analysis algorithm is basically a query expansion algorithm that draws the source of evidence about relevance from the click-through data and can intuitively be explained as follows. This is actually the focus expressed by Li et al. that is most significant to this survey.

A click graph has two types of nodes: query nodes (q) and URL nodes (u). An edge exists between a q and an u when a click has been observed from the click-through data set. After some manual or semi-automatic assessment, some qs represent seed queries and are labeled either as a positive example or negative example. This label information is propagated from seed queries to us and then to unlabeled qs. If a q is a positive example, the linked us become positive examples. If a q is a negative example, the linked us become negative examples.

The efficiency of link analysis algorithms is made high thanks to high matrix sparsity, thus allowing the designer to use inverted file-based indices or data structures for sparse matrices.

Regularization is the another relevant aspect. The click graph that is utilized in the above described algorithm can be sparse since edges between qs and us may be missed because the user did not click on a relevant URL or clicked irrelevant URLs. Missing relevant edges and

occurring irrelevant edges hamper classification. Regularization aims to compensate for the sparsity and noise of a click graph and has been implemented as the posterior probability that a query has an intent.

Markov chains are an alternative, yet similar approach to contextual advertising as suggested by Li et al. [115]. These probabilistic models fit quite well the need of describing the users' click-through behavior from a WWW page to an advertisement without being constrained to the content matching a query (if any). Li et al. believe the user's behavior really reflects which ads may be clicked in the next step. This assumption resembles the assumption made in PageRank such that the authority of the WWW pages depends on the number of paths leading to the pages. The basic idea of Li et al. is that the greater the number of users who clicked an advertisement from a page is, the higher the relevance between the advertisement and the page. Using Markov chains, nodes correspond to the pages and the ads, and edges correspond to the clicks from a page to an advertisement. Edges are weighted so that the higher the number of clicks on an ad and the lower the number of users who clicked the ad, the higher the weight; this is a sort of TFIDF.

A similar yet independently conceived approach to exploiting interaction variables for query intent detection is based on click-through data and session data as illustrated by Cao et al. [36] who have addressed query suggestion. A session is the sequence of queries issued by the same user immediately before the current query; this is a simple instance of context used for detecting the underlying intent. In the authors' idea the intent is represented by the candidate query suggestions. Basically, the authors' approach aims at first observing the context of the current query and then predicting the next queries.

Query suggestion, however, requires clustering individual queries into small groups of similar queries — small groups are necessary to display the queries into small lists displayed below the search boxes, while similarity is necessary to emphasize both common words and the variations between the queries. Clustering queries is however affected by the sparseness of query words. To reduce sparseness Cao et al. enrich the query descriptions with the URLs clicked for queries. In

other words, these authors cluster queries in a click-through bipartite as Li et al. [117] did. Suffix trees are used to organize and sort queries — both small size and high similarity reduce the computational costs for building a suffix tree.

A large click-through data set may provide many information about query intents. Attenberg et al. [11] report a detailed description of a large user study which collected a large click-through dataset through a browser plugin of a major search engine. This design choice allowed the authors to observe the users' trails starting from links displayed in a search engine result page until the (inferred) end of the trail. These detailed data about the users' trails highlight some facts about the relationship between click-through data and query intent. The amount of activity during the trails decreases as the number of terms in the query grows.

Non-navigational queries tend to lead to more click activity than navigational queries, since users tend to move their effort for expressing their own task from the query content to the click-through as the intent moves from non-navigational to navigational. However, this holds when searching for products or services since non-navigational queries that intend to eventually buy something are often more exploratory and a user may not know exactly what is wanted or may require some orientation when searching for a product or service — click-through should provide such an orientation.

The effectiveness of click-through data depends on the amount of historical data which are available for estimation and prediction. This is observed by Shen et al. [156] where it is reported that the performance improvement is more substantial for precision at the top 20 documents than for precision at the top 10 documents. The authors' explanation for this difference between precision at ten and precision at twenty is that after twenty browsed documents the user accumulates more interaction with the system, thus letting the system to calculate better estimations than after ten documents.

2.3.2 Using Implicit Relevance Feedback Data

When it is known, for example by experimental design, that queries have an intent, the relationship between intent and implicit relevance

feedback data about the user's behavior such as document retention and display time is worth investigation.

When attention is paid to the "simpler" query intent rather than to the "more complex" task, computing the data referred to the user's behavior is the main approach adopted within the information seeking and retrieval community. In this respect, Kelly and Belkin [96] found that (i) there is a great variation between subjects in the relationship of display time and usefulness rating, thus making average display time useless, but (ii) there are large variations of display time according to task, thus suggesting that task might be one factor that explains the variability of display time. Unfortunately, there are no consistent findings about the effectiveness of implicit relevance feedback data other than display time.

The idea underlying the use of eye-tracking is that the user looks at the data representing his intent if this is displayed on the screen. An approach to query intent detection using eye-tracking is described by Guo and Agichtein [66]. An indicator of query intent would be gaze position which normally requires eye tracking equipment.

There is a correlation between mouse pointer position and eye position. Therefore, it may be possible to predict the precise times when mouse and gaze position are closely coordinated based only on mouse position and movement. Through a user study, Guo and Agichtein can effectively predict the regions of the screen where the eye and mouse position are within 100 pixels of each other. It is not obvious that mouse pointer position can be detected on a large scale and if it was, it could become an unreliable source of evidence as are most of the sources of evidence that are placed on the client-side. The problem addressed is not at all trivial because a query string is really a very poor source of evidence for predicting query intent and any other source of evidence has the potential to increase the prediction power.

It is possible to predict query intent by looking at the user's past search behavior according to Teevan et al. [163]. To this end, the authors automatically identified a set of navigational queries from the query logs followed by the same result — this identification is based on click entropy. Teevan et al., however, had to make quite a strong yet acceptable assumption, that is, low click entropy is a good approximation of similar intents.

2.4 Detecting Query Intents Using Content Variables

We mentioned that interaction variables such as click-through data, gaze position and mouse pointer position are not always effective predictors of relevance, hence one may put the inappropriateness of these sources of evidence for predicting query intents forward as a hypothesis and investigate whether alternative sources of evidence are more robust and reliable. Content is one of these sources. The content that can be employed to detect query intents derives from documents, queries, search engine result pages, taxonomies or user tags.

2.4.1 Using Document Content

Using only document content for detecting query intents is not frequent — some additional information, perhaps expressed as meta-data are often necessary. Freund et al. [62] associate documents with genre and task, and identify a means of mining various sources of evidence to extract this relationship within a specific domain.

Another example of content combination is when document content is associated with the search engine result pages in which it occurs — the co-occurrence with some other documents is a source of evidence of the relationships between document contents. Yet another example is provided by classification or clustering: for the former document content is associated with the meta-data describing a class, whereas for the latter document content is associated with the other documents of the same cluster.

Query intent detection that is based on search engine result page can be studied within a contextual advertising perspective. To this end, Ashkan and Clarke [9] contribute with some methods and experimental results. Suppose one is interested in deciding whether a query has a commercial intent and then whether some special content variables, for example, an ad has to be impressed on the user's display. In this section, we are interested in the use of search engine result pages for detecting commercial queries which are those queries with the underlying intent to purchase a product or service. In particular, these authors combine two different settings of features: the click-through data features are combined with the query and search engine result page

features, and the combination of query and search engine result page features are used while no click-through data features are involved. The query features that are used by these authors are query length, presence/absence of URL fragments, number of domains listed among the search engine result pages of which the query string is a substring. The search engine result page features that are used by these authors is the frequency ratio of the terms extracted from the first search engine result page.

Query intent detection that is based on classification has been investigated by Broder et al. [28]. Their approach seems promising since the classification accuracy can be maximized by an appropriate quantity of documents given as input. This accuracy rises as the number of documents in a search engine result page increases, and drops when using too few documents due to too little external knowledge, or when using too many results due to extra noise. The optimal number of search results is around fifty. Within the contextual advertising context, Broder et al. addressed sponsored search mechanisms that place relevant advertisements alongside a search engine result page. Sponsored search mechanisms decide whether an ad is impressed and how the impressed ads are ranked alongside the search engine result page. One may correspond to the decision as to whether to impress an ad with the decision as to whether a query has an intent. An ad is impressed if it is decided that the query has a commercial intent, that is, the intent of a user who wants to buy, sell, lease or exchange. The generalization of this sponsored search mechanism to contextual search relies on the use of an intermediate taxonomy of query types chosen according to the intent to be modeled. Broder et al. use a taxonomy available at a commercial search engine for supporting the decision so that queries are classified to the taxonomy classes and then the ads are classified to the queries — Tunkelang [166] surveys the essential concepts of taxonomy-based IR.

Broder et al. [28] have used search engine result pages to obtain additional information for query intent detection. To this end, the authors employ pseudo relevance feedback and assume the top search results to be relevant to the query. As not many results are equally relevant, the given query is dispatched to a general WWW search engine, the top-ranked documents are selected and the WWW pages indicated

by these top-ranked documents are retrieved — note that a very similar procedure that collect the linked pages is used by [101]. Then, the document classifier classifies the search results into the same taxonomy into which queries are to be classified. The classifier was trained by human editors who populated the taxonomy nodes with labeled examples. As the taxonomy included 6,000 classes, simple and efficient classifiers have to be implemented. Once the classifier has been trained, it can be used for classifying queries. This task can be accomplished by computing the probability that a query q belongs to a class c_j and then selecting the class with the highest probability. Taxonomy and classification can be employed in contextual search as follows. Suppose that a user has an intent a when constructing a query q: for all the concepts of a taxonomy, the user first picks a concept c_j with probability $p(c_j|a)$ and then constructs q with probability $p(q|c_j)$ based on the concept c_j. The background knowledge that is provided by the documents is used and then

$$p(c_j|q) = \sum_{d \in D} p(c_j|q,d)p(d|q)$$

is computed. If one assumes that the probability of a query given a document can be determined without knowing the class of the query, one obtains

$$p(c_j|q) = \sum_{d \in D} p(c_j|d)p(d|q).$$

The second step of the approach of Broder et al. is the classification of the query to an ad. Let a be an ad, q be a query and

$$R(a,q) = \sum_{c_j} w(c_j)s(a,c_j)s(q,c_j)$$

be the measure of relevance of q to a. The ss are score functions and w is a sort of prior taxonomy class probability. If

$$s(c_j,a) = p(c_j|a) \quad s(c_j,q) = p(c_j|q)$$

the score functions can be implemented as

$$R(a,q) = \sum_{c_j} w(c_j)p(c_j|a)p(c_j|q) \tag{2.1}$$

given that q and a are independently generated given a hidden concept c_j. The estimation of $p(c_j|a)$ is provided by the document classifier which already estimates $p(c_j|d)$ — the cited paper seems to suggest to approximate the former with the latter. As for the estimation of $p(c_j|q)$, Broder et al. propose a voting scheme. Suppose the r top-ranked document $d_1, \ldots, d_r$ retrieved against q are relevant according to pseudo relevance feedback and $R(d, q)$ is a measure of relevance of d to q. The intuition is that the more the pseudo-relevant documents are likely assigned to c_j, the more the query that matched those documents are likely assigned to c_j. If (2.1) is an optimal document ranking function, top results ranked by a search engine should also be ranked high by (2.1). It follows that when finding the $p(c_j|q)$s, (2.1) must be high when d is a top-ranked document and low when d is a random document. To obtain this result, Broder et al. note that, if $\sum_{c_j} p(c_j|q)^2$ is small, then (2.1) is small for a random document. Therefore, it is sufficient to constrain $\sum_{c_j} p(c_j|q)^2$ to small values when maximizing (2.1). The maximization problem is thus stated as a support vector machine-based classification method, that is,

$$\max_{p(C_1|q),\ldots,p(C_r|q)} \left(\sum_i \sum_{c_j} p(c_j|d_i) p(c_j|q) - \frac{1}{2} \sum_{c_j} p(c_j|q)^2 \right),$$

where the $w(c_j)$s are uniform by assumption.

The approach to detecting query intent based on a large query classification was later extended by a relevance feedback-based method and reported by Broder et al. [30] who basically expanded the feature space as follows. The keywords that occur within the search engine result pages are collected, weighted, and then ranked. The most representative keywords have been used for query expansion. Moreover, the pages returned in the search engine result page are classified to the large query classification. Finally, the set of phrases detected in the search engine result pages and pre-built-in in a lexicon have been prepared. The increase of performance is significantly high.

2.4.2 Using Query Content

The use of query content for query intent detection is grounded on the following associations: A variation of context causes a variation of query

intent which in turn causes a variation of the query. Thus, if different queries are observed and these differences are assigned to intents, a variation of context can be inferred. The problem is, therefore, the assignment of a query to an intent.

Query type-dependent loss functions for training or testing yields better performance than using query type-independent loss functions as showed by Bian et al. Although the numerical differences are small, they are statistically significant. Their approach is described in the following. Suppose that query difference is associated with the user's different expectation on the ranks of the relevant pages in a search engine result page. Suppose also that the simple taxonomy proposed by Broder is considered — it describes query difference based on the search intent of users and classifies queries into intents, that is, navigational, informational, and transactional.

Bian et al. well exemplify how the user's expectations are associated with the way the pages are ranked. Navigational and transactional queries should rank relevant documents on the top of a search engine result page, informational queries should rank relevant documents within the top, say ten ranked documents, multitopic queries should rank some instances of the documents, each relevant to one of the facets, within the top ranked documents, topic distillation queries should focus on ranking a set of documents best representing one single topic among top-ranked documents. Therefore, a query may be assigned to an intent and the intent is a signal of context. However, there are two issues: assign a query to an intent and produce a query taxonomy. This twofold problem can be addressed within a unified decision theory framework as described in the following. Bian et al. [23] develop query-dependent loss functions exploring this kind of query taxonomies. As one of the issues of a taxonomy-based approach to query categorization is indeed the availability of such a taxonomy, these authors further describe a method that learns both ranking functions and query taxonomy simultaneously.

To find differences between queries, a system can be trained after showing it some examples and teaching it to choose the best ranking on the basis of query-dependent loss functions in the learning process — the more the system fails in choosing the document ranking to answer

a query, the higher the loss. The loss function is defined as the sum of the loss functions of a query, the sum being computed over all the query types and over all the queries of a given type — the loss function of a query type may differ for every type. As a risk is an expected loss and the risk implies a probability, probability estimation and in this paper the probability that a query belongs to a query type is necessary. As the estimation of the probability distribution needed to calculate a loss function is difficult, a query categorization would be useful.

Bian et al. propose an algorithm which simultaneously learns a function and categorizes queries. Similarly to the Expectation-Maximization the algorithm is iterative and relies on a mutual relationship. As input, a set of training examples for learning to rank and queries defined by query features are given. As output, parameter vector of the ranking function, ω, and parameter vector of the query categorization, γ, which minimize the loss function, are given. The algorithm starts with random values for ω and γ; it iterates two learning steps until the loss function converges:

(1) the current ω is fixed and the new γ that minimizes the loss function is computed,
(2) the new γ is fixed and the new ω that minimizes the loss function is computed.

Such an approach is feasible and the algorithms converge if loss functions and probability distributions have an appropriate convex form. These details are explained in the cited paper.

Another interesting use of queries is suggested within the context of contextual advertising. Ganti et al. [64] use the corpus of advertising bids used in sponsored search. In sponsored search, each advertiser lists the queries against which an ad should be shown — as this is actually a bid, these queries are called bid-phrases. At retrieval-time, these bid-phrases are matched against an incoming query to determine which ads to display. The advertisers that have listed a bid-phrase matching a specific query is a signal to infer a commercial query intent. The authors leverage the corpus of bid phrases as a set of documents, where each document is "tagged" with the advertiser who has submitted the bid. As each advertiser is interested in listing the

best bid-phrases that match the incoming queries with commercial intent for the products/services it offers, a contextual search system would be concerned in listing the best contextual queries that match the incoming queries with the intent for the actual context.

An immediate application of query intent prediction is to suggest queries to the user. The aim is to predict users' tasks based on implicit relevance feedback data (e.g., user behavior). This problem is addressed by Cheng et al. [44] where the authors propose to mine the latent search intent by using their own framework (i.e., SearchTrigger, that is, a query is triggered by the content of the browsed page) to suggest queries to users when they are browsing. The methodology presented by Cheng et al. is to: produce a series of patterns by using the implicit relevance feedback data observed when the user is accessing a page; extract a series of features from the page and the candidate queries; learn a model for ranking the candidate queries.

A quite elaborate statistical model is crucial at the last step of the methodology because about two thirds of the candidate queries are not of the kind of queries that is useful for the user. Such a statistical model requires that the features are extracted from the queries issued, and from the pages previously visited, by the user. After pages and queries are organized in a bipartite graph, link analysis techniques are applied to assess the popularity of the queries.

If not only the current query but also the past queries are used, it is possible to infer the reasons that lead the user to issue another query. This is a sort of past query intent detection, which is addressed by Cheng et al. [44] who, however, designed their approach to work with user interaction logs storing an anonymous user identifier. As past query intent detection is in this situation tailored to the user, this approach is discussed in Section 3.

Another source of evidence is provided by the time series of the past system's responses, that is, all the responses (up to a maximum time span) are exploited to predict the next response, which has to be decided. Thus, the system's goal is to choose an optimal response when the last action has been observed. So far there is nothing personalized because the user's actions are not labeled by the user. Instead of labeling every user action, Shen et al. [156] propose injecting a user's

mathematical model chosen from a fixed set of user's models. The goal is to minimize a risk function of the user's actions, the system's past responses and the user's model. According to Bayes' rule, the optimal response at a given moment in time is to choose a response that minimizes the risk function. The risk function is the expected loss over the set of the possible user's models. The loss is a function of a user's model, a system's response, the user's current and past actions, and the system's past responses. The probability distribution used to compute the risk is the posterior probability of a user model given all the observations about the user we have made up to time t. A user's model represents what the system knows about the user, for example, the user's information need representation, that is, a bag of words and the documents that the user has already viewed.

2.4.3 Using Document Genre and Relationships

The power of predicting the relevance of the current documents intent may be quite high when the workplace of the users of an IR system is the domains where the effect of the past activities is significant. If a query intent exists, this can be detected by looking at the relationships between documents perhaps along a temporal axis such as the past usage or interactions of the user with the documents. An example of how to study the contextual factors that are related to the user's tasks in a workplace is reported by Freund et al. [62]. The authors conducted interviews with the software engineering consultants of the company used as workplace to understand which factors influence the way the workers search for and select information. Although a workplace seems a limited domain, the documents used by Freund et al. for their investigation are widely dispersed on the company intranet and on the WWW. Freund et al. further investigate the relationship between the user's tasks and document genres. The document genres range from general (e.g., tutorials and presentations) to specific (e.g., technical manuals), thus documents can be categorized by genre. The final goal was to predict task through document genre, and vice versa. An intellectual capital repository of documents either recommended or authored by the consultants has been made available and annotated

with metadata that play the role of feature used in correlating genre with task. Freund et al. state that the existence of a task-genre relationship may help predict task through document genre and that the frequency of co-occurrence of genre features with task features is indicative of patterns of association between genre and task; for example, if users tend to look at a given document genre, a contextual search system may infer the related task. Such an approach has the advantage of requiring that the user knows only which task they want to accomplish. It is useful to note that this investigation was carried out in an enterprise domain and that only half of the queries had a task within the WWW domain as we noted in this section.

An analysis in workplace is also performed by Campbell et al. [34] who argue that IR systems can support the user to accomplish his task by looking at the way in which the documents that are used by the user at a given time had been used at previous times also by other users. Campbell et al. concentrate on the user's activities which involve relationships between documents that can be captured during the course of normal activities in a workplace. Essentially, their approach is centered on a document usage-based similarity matrix which thus defines the contextual relationships between documents. Two types of document relationships have been considered: an undirected relationship (called Common Utility Dependency) such that two documents are related when used within the user's same task (e.g., two documents accessed during the same search session); and a directed relationship (called Reference Dependency) where one document cites another (e.g., in the bibliography). Moreover, the authors have argued that temporal data provides an easy means of interpretation for understanding the current task because the history items can be inter-connected through a network of potentially associated documents for retrieval tasks related to a given activity. The main finding is the high percentage of non-useful documents, especially those relating to the WWW pages that provide the links to useful documents such as the relevant pages, for example, search results pages, contents pages or intranet homepages.

It is worth noting that the idea of context as document network was introduced early by Belkin et al. [18]. Further studies were performed by researchers in automatic hypertext construction who found

that the effectiveness provided by automatic document link detection quickly decreases as the user clicks on documents after issuing a query as reported by Melucci [124]. The situation described by Campbell et al. is complicated by the requirement that a working environment should exploit a combination of historical data such as the time a document is in view. This finding suggests that the existence of an intent behind a query is not obvious.

2.4.4 Using Taxonomies

The advent of the WWW has led to a simplification of the query intent taxonomies. In the current literature, the taxonomy introduced by Broder [26] has become quite well accepted because it allows researchers to simplify the methods for classifying intents. Broder suggests classifying the queries issued to a search engine as informational, navigational, and transactional. We are corresponding the kinds of query introduced by Broder to three query intents, thus introducing three main types of query intent used in the rest of this section. The intent of an informational query is to acquire documents which contain information relevant to the information need of the user. A navigational query involves reaching a known item on the WWW. This intent has a binary outcome — the item either exists or does not exist. A transactional query involves selling, buying, bidding or auctioning items or services. The related queries aim at finding the starting point of a transaction (e.g., shopping, download, access data-bases). An exhaustive and precise query intent categorization is often unavailable and that illustrated by Broder [26] is too general for some applications — a much large query categorization was later introduced by Broder et al. [28] and employed by Broder et al. [27].

According to Dai et al. [49], the classification of queries into navigational and informational is not the only one possible. In electronic commerce, further understanding of commercial intents is crucial. Commercial intent means that the user who issued the query wants to make an action related to commerce such as purchase, auction, selling. Hence, one may suggest that deciding whether an intent exists behind a query is in fact a necessary step before trying to retrieve

possibly relevant documents. Dai et al. [49] report some examples of the relationships between commercial/non-commercial intent and navigational/informational/transactional intents; "u2 music downloads" is a query with transactional and commercial intent whereas "collide lyrics" is a query with transactional and non-commercial intent. In general, commercial intents are behind queries composed of nouns referred to commerce (e.g., "walmart," "dvd," "price").

The most striking feature of Wikipedia is that it is the world's largest knowledge base compiled by humans. The availability of large and manually generated links makes the use of graph algorithms for propagating intent from seed concepts to other concepts feasible. A good example of how external sources of knowledge (e.g., Wikipedia) can be utilized with statistical models is reported by Hu et al. [76] who find that the additional use of Markov chain with Wikipedia provides more benefit than the benefit provided by the use of Wikipedia only. The combination of Markov chains and Wikipedia is also effective for unseen queries, that is, it is possible to predict the intent of queries for which there is not any training data. Clearly, the authors assume that a concept is a good proxy of intent, which might be a valid assumption although query facets would have been a more appropriate term than query intent. Nevertheless, it is likely that for many queries the list of concepts includes some that are related to tasks or intents (e.g., "travel" concepts may include "taxi").

Using Wikipedia for intent detection requires matching the query with Wikipedia concepts. Concepts and articles are organized according to a two-level architecture: the first level includes articles, the second level includes concepts. While articles are linked to concepts by means of a pertinence relationship, concepts are inter-linked through a hierarchical ontology. If the match between queries and concepts is exact, then intent behind the query corresponds to these concepts. Otherwise, the query is mapped to the most related Wikipedia concepts using the features of the query with the title and description n-grams of the top ranked search result snippets from a search engine — these features are weighted by some conventional IR weighting scheme and ranked according to common retrieval algorithms. Hu et al. assume that a travel intention can be detected by searching the query "travel" in Wikipedia.

The concepts that are retrieved would cover almost all aspects of the "travel" domain, for example, travel agency, travel tips, transportation services, such as airlines and taxis, and accommodation, such as hotels and entertainment venues. From these concepts, some seed examples are manually selected. After the seed examples are selected by human experts, Markov chains are used to iteratively propagate intent probabilities from the seed examples into the other Wikipedia concepts, thus assigning an intent probability to each concept. The evaluation was limited to three general applications: personal name intent, job intent, and travel intent. The usefulness of the three general applications stems from the complexity of the underlying tasks. A query has a personal name intent if the query contains a personal name, for example, "john smith pictures" has personal name intent — the significance of this task is confirmed by the very high (i.e., 30%) proportion of search queries with personal name intent in the query logs. Travel is a complex social activity which requires interactions with various agents and regards a variety of aspects such as accommodation and transportation — the significance of this task is confirmed by the relatively high (i.e., 3–4%) proportion of search queries with travel intent in the query logs. The complexity of the travel task generates quite a large number of keywords describing the numerous aspects related to travel. Moreover, these travel aspects are densely inter-linked, thus allowing to disambiguate the words of an aspect through the words of the other aspects. A slightly less significant (i.e., 0.2% to 0.4%) query intent is job intent. Nevertheless, job finding is a crucial task and is carried out by a large number of end users. A query has job intent if the user is interested in finding job-related information such as employment compensation, employment laws, occupations or job WWW sites.

The methods proposed by Hu et al. require somebody to manually label a set of seed queries in order to find some Wikipedia concepts and categories strongly related to the query intent. To this end, a search query log is necessary. For each intent, the human expert uses some queries. After the concepts are retrieved, the parent concepts are retrieved too and the articles linked to these concepts are collected. For large concepts, the number of articles amount to thousands entries. The quantitative evaluations were carried out in comparison with two

baseline supervised methods which do not make use of any additional content variables — these methods exploit the seed concepts only.

The main outcome of the experiments reported by [76] is that the content variables extracted by the search engine result pages is crucial in achieving a better performance. The significance of this outcome is explained by the difference between precision/recall of the baseline methods and precision/recall of the methods proposed in the paper, and by the relatively low performance of the baseline methods, which in fact make use of Wikipedia concepts — therefore, it is not these concepts that contribute to query intent detection, it is on the contrary the content variables provided by snippets and titles of the retrieved documents.

The other interesting outcome is that graph algorithms performed on the graphs induced by Wikipedia concepts and articles may provide some useful evidence to increase the performance of query intent detection. However, this improvement depends on the quality of the manual annotation performed by human experts whereas the above mentioned addition of content variables extracted from search engine result pages is fully automatic. In summary, standard IR techniques can still provide good service.

2.4.5 Using User Tags

Query length and word sparseness are long studied issues of WWW search engines and in general of IR systems. WWW queries contain on average less than three words and there are millions of words in an IR system vocabulary of index terms. Yet the issue of word sparseness is problematic as word-based classification that relies on words may require very large amounts of training data to produce accurate results. As query intent detection leans on classification, these issues affect this task too.

As it was observed as for query expansion when performed by an IR system, one can effectively design and utilize approaches that expand the queries of which intent has to be detected by first issuing the query against an IR system and subsequently extracting additional words from the search engine result pages. These approaches are indeed

effective in increasing the accuracy of query classification significantly. Although these approaches are effective, they are computationally expensive. While the retrieval of search engine result pages is not in itself really computationally expensive, it is their post-processing that can be very expensive and may be in contrast with the most common user's requirements for an almost-immediate answer from the IR system to the typical queries.

A good compromise between effectiveness and efficiency seems a hard result to achieve. Toward this research direction, Ganti et al. [64] consider the tags associated with the pages retrieved in response to a search query and delivered in search engine result pages. Generally, these tags, which are reminiscent of long used metadata, express document properties such as physical organization, topics, authors or logical layout. When tags are associated with search engine result pages, it is possible, for example, to know the distribution of topics retrieved in response to a query. These authors argue that this allows query intents to be determined. In point of fact, the authors show that using tags allows accuracy to be improved in query intent detection since the number of unique tags is smaller than the number of query words, thus reducing the size and sparsity of the distribution of words in documents and the amount of training data required. For each tag, the authors use the fraction of search engine result page pages described by the tag. To this end, the size of the search engine result page for a query and, for each tag, the number of documents in the result tagged with the tag must be computed. The fraction of search engine result page pages described by the tag is computed for suitably chosen sets of query keywords in order to reduce the amount of required memory and thus the computational cost. At retrieval-time, these fractions are retrieved and used as features. Ganti et al. introduce three kinds of tag. Commercial tags are given to products or services searched through queries with commercial intents such as research, purchase or review. These authors observed for this scenario that products and services that are found in search engine result pages are usually the pages returned in response to a query.

When another scenario, for example, entertainment is considered, tags that refers to, for example, entertainment can be found in

Wikipedia articles or in the concepts or categories assigned to articles — the paper written by Hu et al. [76] is also worth reading for the use of Wikipedia for detecting query intent. Wikipedia tags may very frequently be assigned to queries since the search engines often return Wikipedia articles in response to the user's queries and the users tend to get used to these responses and to reuse the same words. Advertising domain tags are assigned to products and services too, yet the intent is slightly different from commercial intents. The intent is in this case retail, that is, the sale of goods to the public in relatively small quantities for use or consumption rather than for resale. As already mentioned in Section 2.4.4, about 6% percent of logged queries contain retail intent.

The three kinds of tag introduced by Ganti et al. correspond to the three tasks evaluated in their research. The first task is classifying product intent for the product category of consumer electronics — as the corpus, the authors use a 7.2 million document Wikipedia snapshot tagged with the 157 product categories. The second classification is to identify queries that are related to health issues. During evaluation a corpus of a hundred million advertisement bids and the thousand most prolific advertisers were used. Each bid phrase was tagged with the advertiser who submitted this phrase to the corpus. The third task is retail query intent classification. To this end, the authors combine the advertiser tags and the Wikipedia category tasks. Each Wikipedia document was tagged by the one thousand most frequent page categories set for the former and the one thousand most prolific advertisers were used. Classifiers were trained by uni-, bi- and tri-grams extracted from training sets either manually labeled by human annotators or semi-automatically labeled by human and computerized annotators. As for the consumer electronics classification task, lexicon features were added. Basically, a lexicon feature tells if a query n-gram includes a lexicon entry. A lexicon is a collection of clusters of words and phrases. For each task, the accuracy resulting from using both the n-gram classifier and the classifier only based on tag frequency, as well as the accuracy resulting from combining both features, were measured. The combination of the two features was based on classifier output combination. The experiment performed for the consumer electronics task resulted that

the n-gram features achieved a comparable accuracy with the accuracy resulting from the combination with lexica.

The use of tag frequencies resulted in a small improvement in accuracy. The same holds for the other classification tasks. The good news is that classification is robust enough when the size of the training set decreases, that is, the accuracy when using tags decreases less quickly than the accuracy when using lexica or n-grams only. This is quite a significant result since training set construction is manually or semi-automatically performed. Overall, it seems that tags provide some useful evidence especially when training sets are small, as long as these tags can automatically be detected from corpora.

2.5 Detecting Query Intent Using Social Variables

In this section, we briefly describe how social variables and in particular the URLs clicked by other users than the current user while completing a task can be used to detect query intent. To this end, we mention a couple of research works that may guide the reader toward further contributions to this topic. In Section 3.4.3, we described the research work by Jones et al. [91] who have addressed the problem of query suggestion through the social variables. An evaluation of an approach to detecting query intents has been performed by White et al. [171] although they focus on personal interest detection. White et al. set up two methods called QueryDestination and QuerySuggestion and asked thirty-six subjects' to use four contextual search systems for completing exploratory or known-item tasks. QuerySuggestion is similar to what is done by Jones et al. QueryDestination suggests a handful of URLs frequently visited by other users who submitted queries similar to the current one.

What White et al. found is that users tend to prefer one method rather than another depending on the query intent. The findings that have been obtained from their study indicate that subjects tend to prefer QueryDestination for the exploratory tasks and QuerySuggestion for the known-item searches (i.e., navigational query intent). Moreover, these two different methods elicit different feelings: QuerySuggestion was felt effective for the known-item tasks because the task is

usually well-defined whereas QueryDestination was felt effective for the exploratory tasks since this task is felt more complex than known-item tasks, the proposed destinations are felt reliable, and the current user trusts what the other users have suggested for similar queries.

2.6 Detecting Query Intents Using Geographical Variables

The easiest and straightforward way to detect query intents is to ask the user to explicitly tell his intent. This is sometimes the case of Geographical IR when for example the user issues queries with geographical operators or the geographical coordinates detected by the user's terminal are automatically added to the query.

The burden of expressing geographical operators is balanced by the effectiveness of these operators when the IR system returns a search engine result page with many pages relevant to the geographical aspects of the user's information need. An approach that is based on geographical operators is reported by Purves et al. [139] who describe the design, implementation, and evaluation of an IR system that is capable of handling queries in the form of the triplet of <theme><geographical_relationship><location>. The core of this system is the indexing process that is capable of detecting geographical data in the documents. The significant effectiveness gained by this system can be considered an upper-bound for those systems which do not burden the user with additional efforts to express their queries and aim at detecting the implicit geographical query intent.

Despite its effectiveness, explicit triplets may become imprecise, incomplete, or very demanding for the user who may not know the name of the location nor the spatial relationships between sub-region of interests, thus resorting to an indirect description of its location and using an alias of a group of scopes. These issues are investigated by Cardoso and Silva [37] who leverage on a geographical ontology for query expansion. A query that may contain candidate geographical data are expanded through the geographical ontology.

Any geographical query processing such as expansion relies on geographical name recognition which is crucial for understanding whether the intent is to retrieve information about a location. In Section 1.3.2,

we pointed out the similarity between geographical name recognition and natural language processing. If this similarity is further investigated, one can realize that the issues of geographical name recognition can be viewed as a subset of named entity recognition, according to Lieberman and Samet [119], although there are some differences.

Geographical name recognition consists of finding all textual references to geographic locations — recognition is then followed by resolution, that is, assigning latitude and longitude coordinates. Effective recognition is necessary to effective resolution since ineffective recognition causes wrong coordinates to actual locations or "right" coordinates to false locations. For example, many proper names of places are also names of people (e.g., Circo Massimo in Rome and Massimo Melucci in Padua share the name yet the former is a place and the latter is a person).

When coupled with social variables, geographical variables can become very effective in detecting a special query intent, that is, looking for news about an event. The basic idea is that there is a significant correlation between a geographical location and an event. This correlation has been tested and implemented by Abrol an Khan [1] who have proposed a geographical contextual search called TWinner. When the user's query includes a location, recent tweets are retrieved if the population density at the location detected in the user's query and the frequency of tweets per minute at that moment are overall higher than a threshold (i.e., 1). This measure indicates that the topic is popular on Twitter and the query is tagged as a query with news intent, that is, queries issued by users who are looking for new.

Geographical variables are effective enough for predicting when a user will click on a news displayed in a search engine result page. Moreover, the power of prediction depends on the population density of the user's location, the distance of the user from the searched event, and the actual location of the search query (measured by an IP address). These findings have been reported by Hassan et al. [70] who investigated the degree to which geographical variables can be leveraged to detect queries with news intents.

Although the use of IP addresses for inferring actual location may be imprecise due to the presence of proxies, the occurrence of

explicit geographical names instead of implicit IP address-based locations in a query does not imply that the query intent is geographical. Nevertheless, Backstrom et al. [13] estimated that the percentage of queries originated from the same IP address is less than 1%.

Geographical names are necessary yet not sufficient to detect geographical intent, thus requiring other query features. An application of this evidence has been reported by Yi et al. [176] and consists of tagging query words using a sort of part-of-speech tagger. The input queries are then split into a geographical subquery and a non-geographical subquery — the union of the two subqueries is the input query.

Queries with geographical intent are significantly longer than queries without, Gan et al. report. However, there are other query intents characterized by long query length since the user tends to add words for making a query special, moreover, a long query may include different intents. Finally, the median of queries with geographical intent is three words which is not very distant from the median of queries with geographical intent, that is, two words. The second evidence reported by Gan et al. is the distribution of query intent across geographical queries and non-geographical queries. They found that queries with geographical intent are more frequently transactional queries (see Broder et al. [29]) whereas query without geographical intent are more likely navigational or informational queries. They also found that queries with geographical intent are about local services such as tourism, government, real estate, education, business, night life, medical, employment, automotive, whereas queries without geographical intent are about links, people, entertainment or people. This distinction is reflected on word distribution across queries.

2.7 Concluding Remarks and Suggestions

Search task detection is a hard problem. Query intent detection is felt easier than search task detection thanks to the recent studies on query intent detection within WWW search that have allowed researchers to correlate some patterns of data with intents. However, there are situations which make search task detection feasible (e.g., in a workplace)

and situations which make query intent detection more difficult than it is felt (e.g., when the query is very short).

A crucial issue is about whether query intents exist, the other is about the technical aspects of the methods for predicting intents if these exist and have to be detected.

Content-based contextual variables are still necessary and sometimes sufficient to the end of detecting query intents. In particular, queries, if available, are the most important source of evidence in IR for representing query intents since queries often yet not always have an intent. Moreover, if additional sources of content-based contextual variables are available, query expansion does still a good service.

However, two important issues should be considered. First, there are intents without queries, this situation being known as zero-query, query-free or no-query IR, which require the utilization interaction variables. Second, there are queries without intents (for example, automatically generated queries) this situation making any contextual variable useless if not noisy. Thus, before engaging in any automatic query intent detection, these two checks are strongly suggested.

What clearly emerges from the literature on query intent detection is the insufficiency of only using interaction variables, these variables being click-through data or implicit relevance feedback data. In particular, automatically detecting intents only from user behavior is a complex issue, does not yield good results and asks the combination of diverse sources of evidence, if possible. As for the effectiveness of using interaction variables, the reader may consult Table 5.1.

An illustration of the computational costs for building a suffix tree is reported by Cao et al. [36] who draw on query clustering for mining context from query history and click-through data, and by Cao et al. [35] who employ click-through data for query classification. These authors show that click-through data is a good source of evidence for query classification and indicate that conditional random fields are an effective statistical model for contextual search.

Further investigation on the user's search session and tasks has been performed by Kotov et al. [102]. Leveling et al. [112] report that

geographical ontology can provide some improvement when integrated with implicit RF.

As regards the issues of training set construction in the context of query intent detection evaluation, the reader may consult the paper written by Li et al. [117] for the solutions of automatically expanding training datasets, which is quite an important problem whenever datasets are sparse as they often are in contextual search.

Query intent detection is described in Rose and Levinson [147] from an information seeking and retrieval point of view thanks to a well-written introduction, related work and useful citations to the background on user behavior in IR.

The importance of geographical names in query expansion has also been reported by Sanderson and Han [151].

Shanahan [154] is an excellent account of contextual advertising that is oriented to an information seeking and retrieval audience.

The TREC relevance feedback track is a source of information about relevance feedback.

Zhuang et al. [183] confirm that geographical ontology can provide some improvement.

3

Personal Interest

Person: from Latin *persona*, "actor's mask, character
in a play," later "human being."
Interest: from Latin *interesse*, "differ, be important,"
from *inter*-, "between" + *esse* "be."
Oxford Dictionary

3.1 Introduction

The users of an IR system are located in a context, place their queries to the system and hope to receive documents that are relevant to their own information needs. This happens in theory, though in practice IR system offer unpersonalized services and are often unable to adapt search engine result pages to the specific user's needs.

At the roots of unpersonalized IR systems is the exclusive use of a content-based query-document matching function for deciding the relevance of the document to the user's information need. As many of these matching functions give high priority to the documents which match a high number of query words, the final effect is that the search engine result pages may include redundant results without meeting diverse, personalized user's needs.

The adaptation of IR system to personal interests is already known in the literature as personalization. Personalization means to design or produce (something) to meet the end user's individual information need. This emphasis on the user helps explain the difference between query intent, which is addressed in Section 2, and personal interest, which is addressed in this section: intent is a property of a query, interest is a property of a user. The former is related to a task which may be relevant to many users, the latter is related to some user's relevant document aspects.

Personalization is not the only term encountered in the literature of contextual search for denoting the adaptation of a contextual search system to the user. Pitkow et al. were among those researchers who distinguished between contextualization and individualization as the two extremes of a wide range of contextual search methods. By individualization Pitkow et al. [137] meant the "totality of characteristics that distinguishes an individual" such as "user's goals, prior and tacit knowledge, past information-seeking behavior." By contextualization they meant the "interrelated conditions that occur within an activity" represented by "factors like the nature of information available, the information currently being examined, the applications in use, when, and so on."

Within personalization, through query expansion, the user's query is in general modified through addition, removal or reweighting of terms for shrinking the query's meaning and fitting the user's interests. For example, the query "IR," when issued by an IR researcher, might be expanded to "information retrieval," to "spectroscopy" when issued by a chemist, or to Spanish when issued by a Spanish native speaker.

Several methods for query expansion are available for adapting a search engine result page to personal interests. Although query expansion is a means for reranking search engine result pages after some additional feedback has been collected from the user for which the contextual search system attempts a personalization, it is not the best approach to personalization since an expanded query may still match little diversified search engine result pages, yet the latter may be different from the previous one. In contrast, diversity is often

preferred when personalization is a requirement because a diversified search engine result page would capture the most salient facets of a query from the most relevant documents at high ranks.

Query expansion is not the only methodology for implementing personalization, search engine result page reranking is another. When reranking search engine result pages, the same query is issued for all users, but the results are re-ranked using interaction, content, geographical or social variables.

An important problem of implementing personalization is whether and how to model the user. While users are thoroughly studied and plenty of models and results have been achieved in information seeking and retrieval, the same success could not be observed for quite a long time in system-oriented IR. One of the reasons was the lack of large user behavior data sets, query logs, geographical coordinates detectors, or social annotation databases. Some progress has been made, though, especially when research works are accompanied by considerations about the role played by the end user when interacting with the system or when they are carried out by using efficient computer systems managing large amount of data by using statistical models. The research work presented in this section illustrates some approaches to this problem.

Another crucial issue addressed in the literature is related to whether personalization is really necessary or useful — although intuition tells us that this is so, the experiments have shown that this is not always the case and appropriate measures should be utilized according to the experience reported in Section 3.2.

3.2 Is Personal Interest Predictable?

People would like an IR system to retrieve the relevant documents as accurately as possible and in particular to filter the documents which are specifically relevant to the user at a given moment in time — this would be the task accomplished by a contextual search system. The problem is that a contextual search system should predict the queries for which personalization improves the results, and the other queries for which it can actually harm.

Whenever the system proposes personalized results to a user who does not want personalization or proposes generic results to a user desiring personalization, the negative feedback may harm the overall system usability for tasks other than the current one as well as possibly limiting the diffusion to the user community. The issues raised by the appropriateness of personalization are addressed by Luxenburger et al. [122] who aim to select the queries that are expected to benefit from the user's history. To this end, the authors introduce different granularity levels of a user profile and propose language models for modeling the user's tasks. These are obtained by means of a hierarchical clustering of the history sessions. The history sessions consist of subsequently posed queries, the result clicks following the queries and the browsed documents within the same session or the documents browsed independently of the query. The approach proposed by Luxenburger et al. might well be placed within Section 2 too. Moreover, Luxenburger et al. have also investigated the utilization of other contextual variable in contextual search as illustrated in this section.

An approach to measure and then detect whether personal interests exist has been illustrated by Teevan et al. [161] who introduced the notion of potential of personalization. Suppose a search engine result page is the response to a query issued by two users. If both users were asked to rate the pages by assigning gains $G(i)$ to every page i of the search engine result page, it would be likely that these two users give different ratings. After a user has rated the search engine result page, the Normalized Discounted Cumulative Gain (NDCG) can be computed for both users. The NDCG is viewed as a measure of the individual user's satisfaction and is a function of both the gains and the ranks. Indeed, the more the search engine result page is ranked according to the user's ratings, the higher the NDCG and the user's satisfaction. Suppose the ratings that a group of users would assign to every page of the search engine result page is the sum of the individual rating assigned by each individual user of the group to the page — this sum would be viewed as a sort of collective rating. The rationale of the sum is that (i) every user equally contributes to the overall group satisfaction and (ii) if the ratings are treated as independent identically distributed random variables, the sum is a random variable following

the law of large numbers.[1] In this way, inference about the overall rating can be made. After this collective rating has been computed, the search engine result page can now be ranked by this collective rating and NDCG can again be computed for this new ranking but using the collective rating. From the user's point of view, this is not likely to be an optimal ranking. The NDCG that is computed for the ranking using the collective ratings is not greater than the NDCG that is computed using the individual user's ratings and the user's satisfaction does not increase when he is presented with rankings "decided" by his group. The latter fact is a consequence of the fact that ranking a search engine result page by the user's rating is optimal for this user and any other ranking is suboptimal. The difference between the NDCG computed from the individual user's rating and the optimal NDCG computed from the collective rating is called potential of personalization, that is, the increase of NDCG that would be achieved if the search engine result page were ranked by the individual user's rating instead of by the collective rating.

The potential of personalization increases as the number of individual users in a group increases as reported by Teevan et al. This trend can be explained in mathematical terms. Suppose m users rate n pages and let $a_{i,j}$ be the ith user's rate of page j. Without losing generality, $a_{i,1} \geq a_{i,2} \geq \cdots \geq a_{i,n}$. When $m = 1$, individual ranking and collective ranking coincide, that is,

$$\frac{1}{m}\sum_{i=1}^{m} a_{i,1} \geq \frac{1}{m}\sum_{i=1}^{m} a_{i,2} \geq \cdots \geq \frac{1}{m}\sum_{i=1}^{m} a_{i,n}.$$

When adding another user to the group, the collective ranking is the same as the collective ranking induced by the m users if and only if

$$a_{m+1,j} - a_{m+1,h} \geq \frac{1}{m}\sum_{i=1}^{m} a_{i,h} - \frac{1}{m}\sum_{i=1}^{m} a_{i,j} \quad \text{for any pages } j,h \tag{3.1}$$

[1] Suppose that $X_1, X_2, \ldots, X_n$ are independent identically distributed random variables each with mean μ and variance σ^2. Then, if $a > 0$,

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq a\right) \leq \frac{\sigma^2}{na^2}.$$

that is, if the rate given to page j by the new user $m+1$ is greater than the rate given to page j, the ranking will be the same. If the rate given to page j by the new user $m+1$ is less than the rate given to page j, the decrease must not be larger than the term on the right-hand side of (3.1). As a matter of fact, the rating given to a page (e.g., page j) by the user $m+1$ is somehow correlated with the ratings given by the other users of the group. Therefore, if the collective group rating given to a page is relatively low, it is probable that the user's $m+1$ rating will be low. When the users of a group independently rate the pages of each other, the inequality (3.1) will often not hold and the ranking will soon change, thus decreasing the individual NDCG and increasing the potential of personalization. When the users of a group tend to conformly rate the pages, high ratings will be given to highly rated pages, the inequality (3.1) will often hold and the ranking will not change, thus not decreasing the individual NDCG nor the potential of personalization.

The potential of personalization can be used to decide whether personalization may be useful. Suppose a collective rating and an individual rating are available. If the gap is larger than a threshold, that is, the ranking based on the collective rating and the ranking based on the individual rating are significantly different, the current user's interests are significantly different from the group's interests. At this point, personalization may be performed since there is a large potential. If the gap is not significantly large, it is likely that the the user's interests are similar to the group's and personalization is of little worth.

When personal interests have a great variation, a user also has great variation in the documents the user considers relevant. Therefore, if the latter were observed, one might infer that some personalization would be useful.

As explicit relevance assessments are difficult to obtain, interaction variables may instead be used. Teevan et al. have utilized interaction data in the attempt to surrogate the explicit relevance feedback data used in the estimation of inter-rater agreement[2] and then

[2] A rater is an agent providing assessments, judgments, rates, etc. on information objects such as documents, terms, links, ads.

in the estimation of the potential for personalization. Teevan et al. assumed that the queries for which there is great variation in the search engine result pages clicked by a user also have great variation in the documents the user considers relevant.

An interaction measure that is used for measuring variability in the search engine result pages clicked is click entropy. Remember that entropy is a measure of the uncertainty in a probability distribution. In this case, the distribution refers to the set of URLs followed after the user has issued a query. Low entropy means that there are a few URLs very frequently clicked and many others rarely clicked (in other words, low entropy means that predicting the most likely URLs is an easy task). Click entropy has been used to detect query intents. While navigational queries are likely followed by only one click, informational queries are followed by more than one click because the end user would need to consult numerous documents for collecting a sufficient quantity of relevant information.

Although click entropy may be used for measuring the variability of click-through as a proxy variable of relevance, click entropy is not enough for understanding whether it is worth applying personalization. For example, given two possible URLs, a query followed by only one URL by 50% of users and followed by the other URL by the other 50% has the same click entropy as a query where every user clicks on both URLs (Teevan et al. [161, 162]). This is because click entropy measures the uncertainty of a distribution of probability that a URL is clicked which is independent of the individuals who actually click, although the variation of the clicks of the first query is high and the variation of the second query is null. The potential for personalization curves better capture variation than click entropy because the curves for queries with the same click entropy but a different average number of clicks per user have a greater gap than queries with high entropy.

3.3 Understanding Personal Interests Using Interaction Variables

To make personalization worthwhile, some information about the user who is interacting with the system is necessary. This information

is most of the time derived from interaction variables, thus placing personalization next to the vast domain of query expansion and relevance feedback — user profiles are sometimes used yet the issues raised by profile creation and update make these data little practical.

With respect to interaction variables, the experiences with explicit relevance feedback data, implicit relevance feedback data and click-through data are the most frequently reported in the literature and are often employed for search engine result page reranking or for query expansion.

Many research works on contextual search have utilized interaction variables for detecting the best search engine result page ranking for a user. Due to the inexpensive means of collection and the wide availability of data sets, click-through data have been the most utilized interaction variables by researchers in contextual search. Because of this relevance, this contextual variable is addressed in Section 3.3.1 whereas the utilization of the other contextual variables that are used for understanding personal interests is described in the remaining sections.

3.3.1 Using Click-Through Data

A method that helps address personalization is the diversification of the search engine result page. The use of click-through data has often been paired with diversification, therefore, we are treating these two topics within this section.

Diversification can be addressed by algorithms that directly learn a diverse ranking of documents based on click-through data. Thus, the problem with diversification can be stated as choosing the optimum set of k pages for a given user population. This is a hard problem since the enumeration of all the subsets of k pages is required, therefore an approximation is necessary. Most learning algorithms aim at maximizing a total payoff or utility function — this is the approach by Radlinski and Joachims [141] and Radlinski et al. [142]. Radlinski and Joachims [141] introduced an algorithm to learn from the preference judgments between pairs of documents. The judgments were not genuine since the click-through data were used to compile them. Suppose

preference judgments over documents d_i and d_j for a given query q are given as input. These preferences are of the form $d_i \succ_q d_j$, that is, d_i is preferred over d_j given q. Preference judgments are similar to relevance assessments, yet there is a quite crucial feature, that is, preferences are necessarily ternary $(\succ, \prec, \equiv)$ attributes of a document pair whereas assessments are n-ary (binary, at least) attributes of a document.

Preferences and assessments are both dependent on the query. Radlinski and Joachims define a support vector machine for classifying documents according to relevance. Learning the weight vector of the support vector machine is based on the preference judgments compiled from click-through data. Radlinski et al. explain how click-through data can be used to learn rankings by maximizing the probability that any new user will find at least one relevant document high in the ranking. In order to provide this explanation, Radlinski et al. [142] report both an offline algorithm and an online algorithm.

The offline algorithm by Radlinski et al. addresses the problem of learning an optimally diversified search engine result page for one fixed query. Suppose each user i has different relevant pages A_i and receives a different search engine result page B_i to his query — it is assumed that every user receives k pages and the query is the same as that of other users yet they have different A_is. When he is presented with B_i, the user clicks on page j with probability $p_{i,j} \geq 0$. Assume that the pages are mutually exclusive, that is, $\sum_{j=1}^{k} p_{i,j} = 1$ for all i. This probability distribution is a user's profile, which might not be completely known. The system gets a payoff 1 if the user clicks, 0 if he does not. The goal is to maximize the total payoff, summing over all users. Under the (strong) assumption that click is a perfect proxy of relevance, the total payoff that is the number of users who clicked on any result can be interpreted as the user finding at least one potentially relevant page. The event that a user does not click is called abandonment since the user abandoned the search engine result page. For Radlinski et al. abandonment is an important measure because it indicates that users were presented with search results of no potential interest. The offline algorithms proposed by Radlinski et al. reach the maximum with probability over a threshold which depends on the number of iterations. If OPT is the maximal payoff that could be obtained if the click probabilities $p_{i,j}$

were known ahead of time for all users and pages, then $(1 - 1/e)\text{OPT}$ is the best obtainable polynomial time approximation. However, even this approximation would require a high computational cost, thus a further approximation is required. An offline algorithm may quite efficiently perform works as follows. Suppose that $0 < \epsilon < 1 - 1/e$ is set such that $(1 - 1/e - \epsilon)\text{OPT}$ is the wanted polynomial time approximation. Then, the number r of iterations is fixed in advance. Moreover, a threshold probability $1 - \delta$ is fixed. A random set of k pages is generated — "random" means that this set is arbitrarily chosen yet any choice criterion should weighed in order to guarantee algorithm convergence speed to the maximum. For each iteration, every page of the k-page set is replaced by one page of the collection at a time and the k-page set is presented to the user who is asked to click on zero, one or more pages. At a rank j, after nr presentations, the algorithm permanently assigns the page that received the most clicks to j, and moves on to the next rank. At the end of the r iterations, the user was asked to click rnk times and every page is assigned a rank between 1 and k. In the paper it is proven that the total payoff is greater than

$$P\left(\text{total payoff} > (1 - 1/e - \epsilon)\text{OPT} - O\left(\frac{2k^2 \log\left(\frac{2k}{\delta}\right)}{r}\right)\right) \geq 1 - \delta$$

that is, the probability that the total payoff is greater than a function of $\epsilon, \text{OPT}, k, \delta, r$ is not less than $1 - \delta$. When k is small enough, r might not be very large. However, the complexity is still high and this algorithm cannot be used online.

The basic idea behind the online algorithm for diversifying search engine result pages that is reported by Radlinski et al. [142] is that every rank j between 1 and k is assigned to a decision algorithm which chooses the page to be ranked at j from the n pages which have not been ranked at ranks less than j. After all the ranks are assigned, a user considers the k pages and clicks on one at most. If the user clicks on a page actually selected by a decision algorithm, the payoff for the algorithm corresponding to that page is 1. The payoff for the algorithms corresponding to all other selected pages is 0 — the decision algorithm are rewarded independently of each other. Radlinski and Joachims have proved that the expected total payoff after r iterations is not less than

$(1 - 1/e)\text{OPT} - O(k\sqrt{rn \log n})$, where OPT is the maximal payoff that could be obtained.

Overall, the retrieval effectiveness that is measured as a fraction of rankings with at least one relevant page is higher than the effectiveness one would expect if other methods purely based on click-through data were used. However, it is lower than methods based on content variables and/or explicit relevance assessments. First, the retrieval effectiveness of both algorithms is higher than 0.70 only if the users are presented with pages in the order of tens-of-thousand times. After one hundred thousand presentations, the retrieval effectiveness increases yet sublinearly.

The offline algorithm appears more effective than the online algorithm, thus reaching an acceptable effectiveness is possible only at a high computational cost. Moreover, k should be kept small in order to reduce the computational cost, however, in this way, highly relevant results that are not ranked within the k pages might never be clicked, which usually leads to the learned ranking never converging to an optimal ranking or giving suboptimal rankings.

Yue and Joachims [179] also address the problem how the user chooses a document reranking out of two rankings displayed to him. However, they bypass the use of proxy variables such as click-through data and propose an approach based on (implicit) feedback gathered directly from users. Their approach assumes that a user decides between two different rankings on the basis of the overall effectiveness — the user looks at the ranking and perceives the overall quality, for example, in terms of precision or NDCG. From an algorithmic point of view of reranking, the problem is to find the best ranking function for a given user as possible. To this end, given the possible ranking functions, the author's approach aims to find a sequence of comparisons between two functions that eventually ends with a close solution to the optimal ranking function. The authors define a Dueling Bandits Problem: The ranking functions in a search engine are points within a space and the only action is dueling two functions (a single comparison between two functions is independent of all other comparisons). A combination of utility, cost, and exploration functions results in an algorithm that eventually selects a quasi-optimal function. It is suggested the reader refer to their paper for the technical details.

A problem that affects the systems that use click-through data for estimating the user's preferences over search engine result pages is the Matthew effect (i.e., the rich get richer and the poor get poorer, see also Section 4.4.1) in the long run. The Matthew effect has been a problem suffered in the application of link analysis algorithms for studying the WWW phenomena and caused by the pervasiveness of the power-laws.

When the Matthew effect affects contextual search, continuing to show the same personalized information to the same users may lead to user frustration. Similarly, not showing personalized information can hurt search effectiveness. Li et al. [116] have addressed the Matthew effect in contextual search through three Exploration-and-Exploitation algorithms which compute an optimum decision between finding novel interesting items (i.e., additional exploration) and suggesting the current items (i.e., additional exploitation). These Exploration-and-Exploitation algorithms share some concepts and notions of Radlinski et al. illustrated above. With these algorithms, the system selects some items and receives user feedback (click or non-click) as payoff. The goal is to find the optimal item selection sequence that maximizes the total payoff. The first two Exploration-and-Exploitation algorithms aim at learning (i.e., optimizing) a probability ξ. With probability $1 - \xi$, this algorithm chooses the best action based on current knowledge; and with probability ξ, it chooses any other action uniformly. The parameter ξ essentially controls the trade-off between exploitation and exploration and it is the analogous of the damping factor introduced in PageRank proposed by Brin and Page [25]. With the first algorithm, a search engine result page that contains k links is divided into three parts. The first part consists of the top ranked k_1 pages and will be reserved for exploitation. The third part contains links to pages that can participate in exploration. The middle part contains pages that are mixed with the first part and the third part by the algorithm according to a random sampling. The algorithm is quite simple. A Bernoulli variable with parameter ξ is sampled until the third part or the second part are empty. According to the Bernoulli outcome, a page is moved from either the first part or the third part to the second part. Such an approach requires the experimenter to decide the optimal ξ where optimality can only be measured through

post hoc experiments. With the second algorithm, ξ is not fixed and the following steps are iterated r times: ξ is sampled from a predefined distribution of probability which is initialized to the uniform distribution at the beginning and dynamically updated by the algorithm. Then, the previous algorithm is performed with the sampled ξ and the resulting search engine result page is presented to the user who clicks on one out of the presented pages. If the user clicks any page, the probability that a ξ is sampled at the next iteration step is increased and the other probabilities are decreased, thus awarding the value of ξ. The third algorithm is an evolution of the first. At the beginning, the first part contains the k_1 most clicked pages as in the first algorithm. These top pages are always kept without changing their rank. The other pages are randomly sampled with a probability distribution defined upon the click-through rate. The lower click-through rate a page has, the more likely it will be selected.

3.3.2 Using Implicit Relevance Feedback Data

Contextual advertising is placed in this section, which is devoted to the use of interaction variables for understanding personal interests, because contextual advertising can be viewed as a kind of personalization, it concentrates on a user and a search engine trying to catch a client, which is a situation similar to that encountered in contextual search. Thus, contextual advertising methods are relevant to contextual search in order to measure the personal user's interest in the clicked pages.

Attenberg et al. [11] pay a great deal of attention on the user activity performed on the sponsored search advertisements displayed by search engines next to conventional organic search engine result pages. In particular, Attenberg et al. showed how to leverage implicit relevance feedback data for adapting ads to the user's personal interests. Their focus is on identifying patterns in user activity and predict expected on-site actions in future instances.

Given a user, a query, and search engine result page, Attenberg et al. wanted to define models for predicting the activity on the trail originating from this result, conditioned on the fact that the result

is clicked. Using only click-through data, they found that the click-through rate of ads does not have a strong correlation with some implicit relevance feedback data since the initial click on an ad is not followed by an intense interaction. This outcome explained the authors' emphasis on detecting personal interests using implicit relevance feedback data.

The emphasis on personalization given by Attenberg et al. relied on a number of interaction variable types such as the number of queries issued by a user, the average number of clicks per query, the probability a clicked result will be an ad, and the expected position of a user's clicked results. These interaction variables were collected from a navigational toolbar plug-in. Therefore, these data refer to individual users yet they have to be anonymized. The click-through data of organic results, sponsored results and the trails followed across the visited WWW sites have been recorded. The mathematical model proposed by Attenberg et al. starts from some empirical observation about the distribution of click-through trail length. According to the previous literature, it is assumed that the probability distribution of click-through trail length follows a power law. However, the parameter of this power law depends on the cluster which includes the query that initiates the click-through trail. Thus,

$$k_c x^{-\alpha_c} \quad k_c = \left(\sum_{x=0}^{\infty} x^{-\alpha_c} \right)^{-1} \qquad c = 1, \ldots, C$$

that is, the probability that the click-through trail length is x is proportional to $x^{-\alpha_c}$. The prior probability that a query belongs to a cluster c is π_c. The mathematical model is defined so that to estimate the unknown parameter vector $\theta = (\alpha_1, \ldots, \alpha_C, \pi_1, \ldots, \pi_C)$. Using this mathematical model, suppose q is a user's query to be classified in one of the clusters. For each cluster c, the probability $p(q|c)$ that q belongs to c has to be estimated. If an implicit relevance feedback dataset is available, the number $n(x, q)$ of click-through trails of length x originated from a query q can be calculated. Under the assumption that the trails are independently observed, we have that

$$p(q|c) = \prod_{x=1}^{\infty} (k_c x^{-\alpha_c})^{n(x,q)}.$$

Thus, the probability that q is observed is

$$p(q) = \sum_{c=1}^{C} \pi_c p(q|c).$$

If the dataset includes the query set Q, the likelihood of θ is

$$\prod_{q \in Q} p(q)$$

and the log-likelihood of θ becomes

$$\sum_{q \in Q} \log \left(\sum_{c=1}^{C} \pi_c \prod_{x=1}^{\infty} (k_c x^{-\alpha_c})^{n(x,q)} \right).$$

Standard maximum likelihood estimation cannot be employed in this case, therefore, Attenberg et al. propose an Expectation-Maximization algorithm for estimating θ.

The search engine result page retrieved in response to the current information need is likely to be highly useful to estimate the current contextual variable. Yet it is not the only source of evidence.

Another source of evidence is provided by the time series of the past system's responses, that is, all the responses (up to a maximum time span) are exploited to predict the next response, which has to be decided. Thus, the system's goal is to choose an optimal response when the last action has been observed. Shen et al. [156] aim to illustrate how to construct and update a user model based on the implicit feedback information. Such an approach can naturally be modeled by classification and in particular by the minimization of a risk function. According to classification, the optimal response at a given moment in time is to choose a response that minimizes the risk function. The risk function is the expected loss over the set of the possible user's models. The loss is a function of a user's model, a system's response, the user's current and past actions, and the system's past responses. The probability distribution used to compute the risk is the posterior probability of a user model given all the observations about the user we have made up to a time instant. A user's model represents what the system knows about the user, for example, the user's information need representation, that is, a bag of words and the documents that the user has already

viewed. In such a framework, the observed data are the set of all user's actions while the system's decisions are the set of all possible system responses to a user's action. The user's actions are modeled as a time series and thus every action is labeled by the moment in time at which it was observed (identical actions are distinct if observed at different moments in time). Instead of labeling every user action, Shen et al. propose injecting a user's mathematical model chosen from a fixed set of user's models. The goal is to minimize a risk function of the user's actions, the system's past responses and the user's model.

Melucci and White [129] present a formal framework based on vector spaces that captures multiple aspects of user interaction and allows a new mathematical model of implicit relevance feedback to be developed. The model uses display time, document retention, and interaction events to build a multi-faceted user interest profile. The paper also introduces some definitions with the aim of providing a useable terminology and a language for describing context in a principled manner. For each dimension of context, first, a set of orthogonal vectors is defined such that each orthogonal vector of such a set models one factor of the dimension of context; second, a basis is built for representing a context by selecting one or more factors from each dimension so that a context is modeled by a set of possible contextual factors and one factor refers to one dimension; lastly, documents are matched against a context by computing a function of the distance between the vector and the subspace spanned by the basis such that the closer the vector to the subspace, the more the object is "in the context." To implement these vector spaces, the vectors that represent the contextual factors are computed by the singular value decomposition of the correlation matrix between the contextual variables observed from a set of documents seen by the user during the course of his search. The function of the distance between the document vector and the subspace spanned by the eigenvector is then used as a measure of the distance between the document and the contextual factor.

What is significant from the paper written by Melucci and White is that it presents a single framework to reason about contextual search and to design contextual search systems. The framework illustrated by Melucci and White [129] is generic and can be applied to a range of

contextual search problems. In this section we describe how the framework can be used to implement an implicit RF algorithm that captures personal interests. In the framework, vectors which represent the contextual factors are computed by the singular value decomposition of the correlation matrix between the contextual variables observed from a set of documents seen by the user during the course of his search. As an example, suppose the following six contextual variable (column) vectors have been observed after seeing six (row) documents:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 & 7 & 6 & 7 \\ 2 & 0 & 9 & 7 & 5 & 6 \\ 2 & 0 & 7 & 6 & 4 & 5 \\ 3 & 4 & 8 & 6 & 7 & 7 \\ 4 & 1 & 3 & 6 & 5 & 5 \\ 1 & 28 & 7 & 7 & 5 & 4 \end{pmatrix},$$

where the columns corresponds to, say, (1) display time, (2) scrolling, (3) saving, (4) bookmarking, (5) access frequency and (6) webpage depth,[3] respectively — all of these values may refer, for example, to time or frequencies, and can be seen as contextual variables of user behavior, which is considered a dimension of context.[4] The following contextual variable correlation matrix is then computed:

$$\mathbf{S} = \begin{pmatrix} 1.00 & -0.42 & -0.14 & -0.78 & 0.11 & 0.05 \\ -0.42 & 1.00 & 0.19 & 0.38 & -0.05 & -0.62 \\ -0.14 & 0.19 & 1.00 & 0.07 & -0.03 & -0.04 \\ -0.78 & 0.38 & 0.07 & 1.00 & 0.00 & 0.00 \\ 0.11 & -0.05 & -0.03 & 0.00 & 1.00 & 0.75 \\ 0.05 & -0.62 & -0.04 & 0.00 & 0.75 & 1.00 \end{pmatrix}.$$

The values of an eigenvector of $\mathbf{S}$ are scalars between -1 and $+1$; the further a value is from 0 the more the value corresponds is a significant descriptor of the contextual factor represented by the eigenvector. The value can be likened to an index term weight. The sign can express the contrast between contextual variables

[3] The depth of a webpage is the number of links from the root of the web site to the webpage itself.

[4] This example is reported in the Melucci and White [129] paper.

and then the presence of subgroups of contextual variables in the same contextual factor. For example, the first eigenvector is $\mathbf{b}'_1 = (-0.479, 0.516, 0.170, 0.436, -0.308, -0.436)$ and it tells us that saving is of little importance, since $|b_{i3}| = 0.170$ is relatively close to zero, while the most important contextual variables tend to cluster: scrolling and bookmarking tend to be performed together ($b_{i2} = 0.516, b_{i4} = 0.436$) and tend not to be performed when display time, access frequency, and browsing ($b_{i1} = -0.479, b_{i4} = -0.308, b_{i6} = -0.436$) increase. Let $\mathbf{b}_i$ be one of these eigenvectors and $\mathbf{y}$ be an unseen document. The function of the distance between the document vector and the subspace spanned by the eigenvector is then used as a measure of the distance between the document and the contextual factor. Therefore, $\mathbf{y}' \cdot \mathbf{B}_i \cdot \mathbf{y}$ is computed. If the unseen document vector is, say, $\mathbf{y}' = (0.71, 0, 0, 0, 0.71, 0)$, then the distance is 0.31.

The framework presented by Melucci and White [129] has been summarized in Table 3.1. Although this is a general framework, we decided to summarize it in this section, which is devoted to personal interests, and not in a higher section since the survey is about the literature on contextual search and not on the specific author's point of view of contextual search. In this view, context is a vector space and every (e.g., document, query, factor, variable) is a vector. A vector space is defined as a coordinate system whose canonical (or standard) basis corresponds to the contextual variable — every vector of the space is defined with respect to the canonical bases. Beside the canonical basis in this vector space, many vector bases can be defined. Every vector basis that is not the canonical basis refers to either a contextual factor or a contextual variable. A vector that represents an information object such as a document or a query is generated by a vector basis in the same way as the information object is generated by a contextual factor

Table 3.1. A vector space that represents the contextual factors and contextual variables described in this survey.

Concept of this Survey	Vector Space Concept
Context	Vector space and its canonical basis
Contextual factor	Vector basis
Contextual variable	Vector basis

or a contextual variable. Every vector can be generated by a different basis in the same way as an information object is generated in different contexts. The probability that an information object is in a context is a function of the projection of the information object vector on to the subspace spanned by the basis vector corresponding to the contextual factor. For example, consider a three-dimension vector space defined over the real field. The canonical vector basis is then

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Suppose there are two contextual variables: click-through rate and term occurrence. The former has values over the field, that is, $\{0, 1, n\}$, where n means "two or more," and is represented by the vector basis

$$\begin{aligned} \mathbf{c}_0 &= c_{0,1}\mathbf{e}_1 + c_{0,2}\mathbf{e}_2 + c_{0,3}\mathbf{e}_3 \\ \mathbf{c}_1 &= c_{1,1}\mathbf{e}_1 + c_{1,2}\mathbf{e}_2 + c_{1,3}\mathbf{e}_3 \\ \mathbf{c}_n &= c_{n,1}\mathbf{e}_1 + c_{n,2}\mathbf{e}_2 + c_{n,3}\mathbf{e}_3. \end{aligned}$$

The latter has values over the set $\{0, 1\}$ (i.e., presence, absence) and is represented by the vector basis

$$\mathbf{t}_0 = t_{0,1}\mathbf{e}_1 + t_{0,2}\mathbf{e}_2 + t_{0,3}\mathbf{e}_3 \quad \mathbf{t}_1 = t_{1,1}\mathbf{e}_1 + t_{1,2}\mathbf{e}_2 + t_{1,3}\mathbf{e}_3.$$

Suppose there are two contextual factors: query intent and personal interest. The former has values over the field, that is, {navigational, transational, informational} and is represented by the vector basis

$$\begin{aligned} \mathbf{q}_1 &= q_{1,1}\mathbf{e}_1 + q_{1,2}\mathbf{e}_2 + q_{1,3}\mathbf{e}_3 \\ \mathbf{q}_2 &= q_{2,1}\mathbf{e}_1 + q_{2,2}\mathbf{e}_2 + q_{2,3}\mathbf{e}_3 \\ \mathbf{q}_3 &= q_{3,1}\mathbf{e}_1 + q_{3,2}\mathbf{e}_2 + q_{3,3}\mathbf{e}_3. \end{aligned}$$

The latter has values over the set $\{0, 1\}$ (i.e., not interesting, interesting) and is represented by the vector basis

$$\mathbf{p}_0 = p_{0,1}\mathbf{e}_1 + p_{0,2}\mathbf{e}_2 + p_{0,3}\mathbf{e}_3 \quad \mathbf{p}_1 = p_{1,1}\mathbf{e}_1 + p_{1,2}\mathbf{e}_2 + p_{1,3}\mathbf{e}_3.$$

Through easy algebraic manipulation, one can show that the contextual factors can be expressed as linear combinations of the

contextual variables, thus representing the relationship between them within a single vector space. The probability that the query intent is navigational when the click-through rate is two or more is expressed as

$$p(q_1|c_n) = |\mathbf{c}'_n \mathbf{q}_1|^2.$$

The probability that the query intent is navigational when the click-through rate is one or more is expressed as

$$p(q_1|c_1 \vee c_n) = \mathbf{q}'_1(\mathbf{c}_1\mathbf{c}'_1 + \mathbf{c}_n\mathbf{c}'_n)\mathbf{q}_1 = |\mathbf{c}'_1\mathbf{q}_1|^2 + |\mathbf{c}'_n\mathbf{q}_1|^2$$

(see also Melucci [125]). As the canonical vector basis are orthonormal, we have that

$$p(q_1|c_n) = |c_{n,1}|^2|q_{1,1}|^2 + |c_{n,2}|^2|q_{1,2}|^2 + |c_{n,3}|^2|q_{1,3}|^2$$

and that

$$\begin{aligned} p(q_1|c_1 \vee c_n) = {} & (|c_{n,1}|^2 + |c_{1,1}|^2)|q_{1,1}|^2 + (|c_{n,2}|^2 + |c_{1,2}|^2)|q_{1,2}|^2 \\ & + (|c_{n,3}|^2 + |c_{1,3}|^2)|q_{1,3}|^2, \end{aligned}$$

where $|c_{i,1}|^2 + |c_{i,2}|^2 + |c_{i,3}|^2 = 1$. This equality similarly holds for the qs and the ps.

In the rest of this section, we explore the work of Joachims et al. in greater detail since it provides a machine learning-based approach to understanding personal interests and in general relevance assessments using implicit relevance feedback data. This approach to ranking documents can be extended to personalization since the queries can be labeled by user labels if appropriate interaction variables have been observed.

One of the most significant drawbacks in using click-through data is the bias caused by the system's ranking and presentation style. This bias makes the inference from the user's click-through data almost impossible or at least unreliable. Joachims et al. found that users still click on the top-ranked URLs of a search engine result page even when the most relevant URLs are placed at the very bottom of the search engine result page. A reasonable explanation of the user's rigidity in choosing the top-ranked URLs and in sequentially scanning (see Joachims et al. [88]) the results from the top to the bottom may be

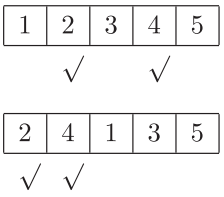

Fig. 3.1

based on the same assumption which is underlying every IR system, that is, such a system is designed to serve users as best as possible and users trust such a system. Nevertheless, this explanation contrasts with the fact that diverse users have diverse and personalized needs, thus a one-size-fits-all IR is unlikely to serve equally every user. What Joachims et al. have done is to concentrate on unclicked URLs and not only on those clicked. Indeed, what really matters is to know whether and possibly why a user did click a result and not click another result. Therefore, the clicked page is *preferred* to the unclicked page. This view is compared with the traditional way of collecting relevance assessments through the example of Figure 3.1 which shows a five-page search engine result page and the clicked pages.

An immediate advantage of viewing judgments as preferences is that a single click or other relevance judgment over a result "spreads" a high number of preferences since the clicked result becomes preferable to any other — from one assessment to one million assessments. Preferences allow effectiveness measures other than traditional precision-based measures to be utilized. To this end, consider the Cartesian product between the set of retrieved documents and itself. A search engine result page (or ranking) can be viewed as a subset R of this Cartesian product such that a document pair (d_i, d_j) belongs to R if and only if $d_i \prec d_j$, where $\prec$ means "is ranked before." Given two rankings R_1, R_2, a measure of correlation is Kendall's τ and the correlation is expressed as $\tau(R_1, R_2)$. Kendall's τ is a function of the number of exchanges of documents necessary to transform one ranking to the other ranking. The idea underlying the use of τ is that of independent arrangement: If we observe two variables as they were observed from two seemingly independent urns, and if the values of a variable

are extracted in increasing order, the extent to which the corresponding values of the other variable depart from the order indicates the weakness of the correlation between the variables. In the example, $D = 3$ pairs are discordant (i.e., $2,1$, $4,1$, $4,3$ of the second ranking disagree with $1,2$, $1,4$, $3,4$ of the first ranking) and $C = 7$ are concordant. Thus, $\tau = (C - D)/(C + D) = (7 - 3)/(7 + 3) = 4/10$. A recent illustration of τ in IR evaluation is given by Melucci [126].

Joachims [86] has argued that τ is appropriate for IR since it is related to other measures. We add that τ is connected to Binary Preference introduced by Buckley and Voorhees [32] as an alternative to the classical retrieval effectiveness measures for facing the problem of incomplete judgment. BPref is the average degree to which *judged* non-relevant documents are ranked before a relevant document computed over all the relevant documents. BPref was defined as

$$\frac{1}{\text{rel}} \sum_{r=1}^{\text{rel}} \left(1 - \frac{|\text{non-relevant documents ranked before } r|}{\min\{\text{rel}, \text{nrel}\}} \right),$$

where rel is the number of relevant documents and nrel is the number of non-relevant documents. It can be shown that $\text{BPref} = C/(C + D)$. (See also the textbook written by Croft et al. [48, Ch. 8].) It follows that $\tau = 2\text{BPref} - 1$. Suppose the queries of a training set of n examples have been associated with their *optimal* search engine result page (i.e., ranking) $R_1^*, \ldots, R_n^*$, have been given to the retrieval function of an IR system as input, and have been answered by a ranking R_q for every example q. However, an IR system cannot produce an optimal ranking for each query. Therefore, the problem is to learn the retrieval function that maximizes the average empirical τ between a ranking and the training examples, this τ being defined as

$$\tau(f) = \frac{1}{n} \sum_{q=1}^{n} \tau(R_q^*, R_q).$$

Consider a training set containing the five documents of Figure 3.1, three instances of the same query q, and an optimal ranking for each instance of the same query; for example, the training set may appear

as follows:

1	2	3	4	5
2	4	1	3	5
3	5	1	2	4

If another instance of the same query is given as input, the retrieval function has to find the best ranking. To this end, a value of τ is computed between each example of the training example and each permutation of the five documents, thus resulting in $3 \times 5! = 360$ calculations. A trivial and naïve approach is then based on enumerating and computing $\tau(f)$ for all the possible permutations R_q. However, this approach is impractical due to the factorial number of permutations.

The second problem is that a query that has no example in the training set cannot be answered. This is a consequence of the natural phenomenon that there are very few frequently observed queries and many infrequently observed queries, as well as many new unobserved queries. Instead of learning a function that computes the best permutation (i.e., ordering) of the documents, Joachims et al. have introduced a utility function that assigns a real-valued utility score to each document d and query q. The basic idea is that whenever a document d_i is preferred to document d_j for a query q, then the utility perceived by the user who issued q and computed for d_i is greater than the utility computed for d_j. To find the utility function, for each query q, every document is mapped to a real vector of contextual variables — a contextual variable is then indexed by both a document and a query. As the input query q and all the documents $d_1, \ldots, d_i, \ldots, d_j, \ldots$ in the training set are known, every contextual variable vector $\mathbf{v}_q(d_j)$ is known. Learning the utility function is then learning a real parameter vector $\mathbf{w}$ such that $\mathbf{w}'\mathbf{v}_q(d_j) > \mathbf{w}'\mathbf{v}_q(d_i)$, whenever d_j should precede d_i in the ranking produced by the utility function. Consider the training set containing four documents of Figure 3.1, three queries and four documents. For each document-query pair there is a real contextual variable

vector of $\mathbb{R}^5$, for example:

Document	Contextual Variable Vectors				
Rank	1	2	3	4	5
		query 1			
3	1	1	0	0.2	0
2	0	0	1	0.1	1
1	0	1	0	0.4	0
1	0	0	1	0.3	0
		query 2			
1	0	0	1	0.2	0
2	1	0	1	0.4	0
1	0	0	1	0.1	0
1	0	0	1	0.2	0
		query 3			
2	0	0	1	0.1	1
3	1	1	0	0.3	0
4	1	0	0	0.4	1
1	0	1	1	0.5	0

Suppose that the four documents have to be ranked against another query using the following contextual variable vectors:

Document	Contextual Variable Vectors				
Rank	1	2	3	4	5
		new query			
4	1	0	0	0.2	1
3	1	1	0	0.3	0
2	0	0	0	0.2	1
1	0	0	1	0.2	0

where the leftmost column includes the expected document rank. Using the Joachims [87] package, the calculated document ranking exactly matches the expected ranking.

3.4 Understanding Personal Interests Using Content Variables

3.4.1 Using Query Content

Query ambiguity is perhaps the most noticeable symptom of the need for personalization — for decades, it has often been noted in the literature that two users issuing the same query often have different intents and give different meaning to the query words.

Although the findings are not always consistent across different research works, it is useful to highlight the most significant results especially in the light of the design of contextual search systems.

Query expansion is perhaps the most widespread method for extracting evidence about personal interest and in general from context. The paper written by Pitkow et al. [137] was one of the earliest on contextual search and in particular on using query expansion for meeting personal interests. To our knowledge, they were the first to mention the idea of comparing the current query with something else for deciding whether personalization is worth performing. In this respect, Pitkow et al. proposed comparing the query term and a user model representing what the user has seen before. If the query is similar to what the user has previously seen, the system can reinforce the query with similar terms, or suggest results from prior searches. If the query is a new topic, the system should not expand the query or it could help define what the topic is not about by providing a diverse set of results to the user — the latter is the standard semi-automatic query expansion. Interestingly, Pitkow et al. mentioned that user profiles can serve as filters for ranking the search engine result page.

The specific problem of personal navigation queries is addressed by Teevan et al. [163]. A navigational query is personal when there exists one user who more frequently issues the query than other users. A personal navigation query looks like a general informational query since it is shorter than the average query yet longer than general navigation

queries. They are less likely to contain URL fragments and less popular than average queries, yet they make up a high proportion of the queries issued by an individual. As a general result, the authors found that personal navigation queries can be predicted by looking at the number of times the query is issued — the more the user issues the query, the more likely the query is navigational for the user.

3.4.2 Using Search Engine Result Page Content

In Luxenburger et al. [122] the following data were observed and stored to a local database file: the URLs clicked on search engine result pages, HTTP traffic, queries, clickstream of subsequently visited WWW pages. Luxenburger et al. defined task as cluster of user profiles in their work, therefore the sense of "task" appears closer to what we mean as user than to what we mean as task or intent in Section 2. Due to this emphasis given by these authors, their work is described in this section.

The interest in the approach proposed by Luxenburger et al. is due to the clustering of data of diverse types, that is, interaction variables (e.g., click-through data) and content variables (e.g., queries and browsed documents) observed within a search session. Clustering has been hierarchical, thus making it possible to have small tasks consisting only of a single session, to the largest task encompassing the whole user search and browse history. The same hierarchical clustering algorithm is used for clustering search engine result page snippets and titles and obtaining candidate query facets $F_1, \ldots, F_m$. The approach of Luxenburger et al. aims at comparing the tasks (i.e., search sessions) $T_1, \ldots, T_n$ with the facets. To this end the tasks and facets are represented by unigram language models and comparison is measured by the Kullback-Leibler divergence between a query facet F_i and a task T_j. If the Kullback-Leibler divergence is larger than a threshold, that is, task and facet are significantly different, the current user's profile is different from the intended meaning of the query, thus refraining from personalizing the search results. Otherwise, it is likely that the query refers to the user's profile and the slight differences can be reduced and the search engine result page personalized. Personalization is implemented

by query expansion and adding the best discriminating query facet from all other query facets, while being most similar to the task. Luxenburger et al. mix different models and give higher weight to the queries submitted later in the session since they are likely to be better matches to the user information need and likely to better characterize the user's intention than the old queries. Such flexibility is not only useful when designing IR algorithms but also during the experiments when the researcher needs to test and tune different parameters.

An aspect that should be considered when designing personalized ranking methods is efficiency. Indeed, a contextual search system should choose between client-based personalization (user profile representation resides on the client size) and server-based personalization (document representations reside on the server side). Teevan et al. [160] think that the most efficient solution would be to implement the methods entirely on the user's machine, thus only downloading the search engine result pages. According to their criteria, Teevan et al. suggest that the most effective and efficient combination of parameters is: corpus representation approximated by the result set title and snippets, which is inherently query focused; user representation built from the user's entire personal, query focused index; document representation based on the title and snippet returned by the search engine; query representation built by query expansion based on words that occur near the query term.

However, it is both worthwhile and necessary to note that moving all the computational costs on the client-side will result in an inefficient solution if the user's machine has little power (e.g., a standard mobile phone) or if the network bandwidth is not broad enough. Efficiency may be dependent on the user and the query, thus making query performance prediction or user profiling somehow necessary tasks.

3.4.3 Using Profiles and Categories

Predefined categories drawn from some directories or user profiles are a quite frequently investigated approach to personalization. User profiles can give appreciable improvements and that, in particular, personal profiles integrated with general user profiles are more effective than

either a personal or a general user profile alone. Similarly, using categories (e.g., those from the ODP) is useful to improve effectiveness according to Ma et al. [123].

When query expansion selects the number of expansion terms depending on the user and on the the user's query, it outperforms both the original ranking and the personalization in the case of a fixed number of expansion terms as Dang and Croft [50] and Luxenburger et al. [122] report. A formalization of the combination of click-through data, content and user profiles has been described by Sontag et al. [158]. Basically, probability distributions were extensively used in that paper for modeling every entity playing a role in a contextual search system. Thus, relevance and contextual variables are modeled as random variables, feedback is modeled as probability update through the Bayes rule, decision is supported by divergence measures.

When query expansion is insufficient, it might be integrated by the user's search history as proposed by Liu et al. [121] who propose modeling and gathering the user's search history for constructing a user-focussed profile and a general profile based on the ODP and for deducing appropriate categories for each user query based on the user's profile and the general profile.

The combination of general and personal profiles is crucial to reduce the problems due to bias and fitting. Liu et al. relate a search history to every user. A user search history consists of queries, relevant documents, and related categories and is represented as a tree model where nodes are information items and edges are relationships between nodes. The root of a search record is a query. Each query has one or more related categories. Associated with each category is a set of documents, each of which is both relevant to the query and about to the category — in this way both relevance and aboutness are considered in the same model. After building the tree-model of the user search history, a user profile can be built from a set of categories and a set of weighted keywords can be built for each category.

The key notion of the user profile is that each category represents a user interest in that category. The user's interest in the category is tuned by the weight of a keyword. Mathematically the user's search history is represented by a relevance matrix which is constructed from

the user queries and the relevant documents (here a document is either a query or a relevant document), and an aboutness[5] matrix built from the relationships between the categories and the documents. Given a relevance matrix and an aboutness matrix, the aim is to compute a user matrix that represents the user's profile. The matrix that describes a user is the result of the product between a category matrix and a keyword matrix. The rationale behind this product is that a keyword matrix associates the keyword to the categories but this association is filtered by the user's profile because the final aim is to detect keywords which match the user's categories and are good candidates for query expansion. Liu et al. illustrates four different methods which have some matrix operations in common.

The common idea of these methods is that a relevance matrix and an aboutness matrix are related by the user matrix through linear operations. In particular, it is assumed that the relevance matrix, which is known, is the product of the aboutness matrix, which is known too, with the user matrix, which is the unknown. To compute the user matrix, regression is implemented by using singular value decomposition or linear model — the details are in their paper. While a user profile is a set of categories in the paper written by Liu et al. [121], it is a dendogram[6] according to Luxenburger et al. [122]. A dendogram ranges from small sessions consisting of only a single session to the whole user search and browse history.

The idea of structuring the user's profile in a hierarchical structure has been transposed to search engine result pages. Each result is represented by its title and snippet information to obtain candidate query facets which represent the different aspects the query might span. Each obtained query facet and each session is represented by a unigram language model.

As mentioned above the indiscriminate application of personalization, for example, by using a category-based profile as proposed by

[5] The term "relevance" and "aboutness" are adopted by us to stress the combination of relevance and aboutness.

[6] A dendrogram (from Greek dendron "tree," -gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering (from Wikipedia [175]).

Liu et al. [121] to all the user queries, is not always appropriate. Indeed the user may feel disturbed by personalization when accomplishing her task because what has been thought "personal" for one user does not necessarily fit another user or no longer fits the same user even after a short time. Such a feeling happens because user interests change over time: a user sometimes is interested in different categories in the same timespan, or even personalization may impede a user's desire of exploring new topics.

3.5 Understanding Personal Interests Using Social Variables

Social variables are observed from user groups. When social variables are used to understand personal interests, it is crucial to detect the user's groups. The importance of correctly detecting the user's groups has been found by Teevan et al. [164] who investigated the effect of user group profiles on personalized search.

The most important finding is that groups can provide effective evidence for detecting personal interest only if these groups have cohesive interests, otherwise the retrieval effectiveness can significantly drop. This is the reason that social tagging, which is addressed in the remainder of this section, can effectively work: social tags is a means for making groups cohesive since the users of a group share the same tags and are then likely to share the same interests.

The users of social tagging systems organize and share their own and the other users' content. Through tags, users annotate each document with any number of freely chosen words.

The combination of organization, sharing and tagging makes social tagging systems an interesting playground where the social dimension and the personal dimension interact with interesting results.

The main problem for the designers of social tagging systems enhanced with contextual search functionalities is that the simplistic and unrestricted vocabulary is a reason of failure in retrieving relevant documents, yet it is a great facilitator for the users.

A significant result of the interaction between the social dimension and the personal dimension is that the personalized categorizations and annotations defined by a user can in principle help the other users in

locating documents. In this way, personalization can leverage the large size of social variable data sets.

The world of social tagging is addressed by Harvey et al. [69], for example, who address the missing match between annotation vocabulary (i.e., the set of words used by the users for tagging) and user queries due to the data sparsity which is in turn caused by the large size of the vocabularies and the shortness of the queries. In their paper, Harvey et al. describe two applications of the latent Dirichlet allocation models.

A latent Dirichlet allocation model views entities (e.g., documents, words, topics, users) as variables and the observed values of these entities as outcomes of these variables. Not only does each outcome have a probability of being observed, but it also has a series of conditional probabilities, that is, the probability that an outcome is observed conditioned to the observation of an outcome of another variable. Another feature of latent Dirichlet allocation is that the probability distributions of the outcomes are governed by parameters which are in turn variable — the parameters are described by Dirichlet distributions. Moreover, there exist latent variables and observed variables in latent Dirichlet allocation — the latent variables are often called topics.

When applied to social tagging, latent Dirichlet allocation models are used to describe three entities, that is, documents (or resources), users and tags where each entity is described as a vector of words from a common vocabulary. Tags, documents, users, topics, and words are assigned a probability distribution defined in terms of the conditional probability distributions of the entities. Defined in this way, latent Dirichlet allocation models make it possible to infer the topics that are in common to the users on the basis of the tags assigned to the documents.

Another use of social variables occurs when the users are interested in some personally selected topics. The problem of tracking and detecting topics in a continuous stream of short natural language texts is addressed by Lin et al. [120]. Although it resembles Information Filtering or Topic Detection, these texts that have to be filtered are actually short (max 140 characters) messages called “tweets” which are sent

by users called "followers." Tweets are a difficult source of evidence because they: are very short; convey little information; change rapidly over time.

Another source of difficulty with "tweets" is due to the fact that the hashtags are not elements of any syntax, but were invented by the followers to describe their tweets. This means that nothing can prevent the advent of some special tag. Moreover, computational problems are quite challenging because of the high arrival rate of tweets. Lastly, although the topics of interest are generally coherent and stable over short time intervals, tweets may be about many different entities.

Despite the computational issues, tweets offer interesting opportunities since they might be a source of evidence about what users are doing (task), who they are interacting with, where they are located, thus making them a potential source of evidence for contextual search. In Lin et al. [120] the authors use language models which are trained using hashtags and are adaptive because new tweets update the probability distribution. The language models then filter the tweet stream by computing perplexity. Perplexity is an Information Theory measure of the degree to which a probability distribution (i.e., a Language Model in this case) can generate a tweet — perplexity is actually quite similar to the statistical likelihood function.

Language models allow Lin et al. to model the evolution of hashtags: the authors want to detect recent tweets yet do not want lose past tweets. A good smoothing of "foreground" and "background" provides effective estimates of term distributions across time. Language models provide a good tradeoff between novelty and sparsity. The results of this work suggests that simple approaches work well in terms of tradeoffs between efficiency and effectiveness.

An interesting situation that emphasizes personal interests and non-textual media is image retrieval. von Ahn and Dabbish [168] introduce a game such that the people who play the game label images for us. Their system, called "the Extra Sensory Perception game," is played by two participants who are not necessarily subjects of a user study. This game is instead intended as an online game played by a large number of players. Participants are assigned images so that for each image there are two or more participants playing with it and each

pair of participants share an image. Each player has to guess which word the other participants is going to type by typing the word. As both participants have in common only the image about which they are typing a word, they tend to type words that are related to the image — the game ends when both participants type the same word, thus agreeing on the content of the image.

The idea underlying the Extra Sensory Perception game is similar to the ideas underlying other approaches described in this section: the mutual relationship of Hyperlinked Induced Topic Search; the endorsement relationship of PageRank; the reuse of hashtags in tweets.

Overall, the results reported by von Ahn and Dabbish indicate that these games are an effective means for describing large quantities of images in a short time. In particular the results indicate that: participants in the Extra Sensory Perception game are willing to play, the latter being not taken for granted; there was a high inter-participant agreement on the words assigned to the images; the majority of the words are useful in describing each image; the majority of subjects think almost no word is unrelated to any image — the reader may want to read Robertson et al. [146].

When personal intents are related to what and where a particular user was interested or paying attention to, large photo collections can be leveraged. Yet, these collections need representation and often interconnection to be effectively used. Tuite et al. [165] introduced a similar "gamification" of the solution to the problem of adding representation to documents. Their PhotoCity is a game that allow players to capture photos and that processes the photos using computer vision techniques for incrementally expand 3D models. The players capture virtual flags and castles, thus contributing to digital 3D replicas of real buildings. The result of the game is a set of detailed 3D models and a large set of photos that densely cover an entire area from many different angles.

In this section, we are also describing another approach to personal interest detection which is based on the past queries issued by other users and is aimed at modifying the current user's query. The approach is by Jones et al. [91] who described a method for query modification that is based on past users' queries, phrase similarity, and query suggestion ranking. Their results are striking and are worth description.

The basic idea is to exploit candidate reformulations expressed as pairs of successive queries issued by a single user on a single day. For each pair (q_1, q_2), it is assumed that q_i is a realization of a Bernoulli random variable with parameter p_i. Then, a hypothesis test is performed with null hypothesis $p_1 = p_2$. As is customary, a log-likelihood ratio is computed. To solve the problem due to very few frequent queries and many rare queries, Jones et al. segment a query into phrases, thus collecting much more statistical evidence for computing the likelihoods.

Another use of social variables for tailoring documents to the personal interests has been described by De Francisci Morales et al. [51]. In that paper, they observe that the "tweets" relevant to a news item precede press release and decrease while newswires are continuing. The social network of micro-blogs can then anticipate the news relevant to the user's interests. The model proposed by De Francisci Morales et al. recalls the vector space model since it is a mixture of matrices connecting users, "tweets" and news.

Finally, a combination of social variables and geographical variables is described by Kinsella et al. [100]. This is another example of how language models can be exploited for modeling and integrating diverse contextual variables together. In these contexts, it is crucial to define the best contextual variables and then to accurately estimate the probabilities as possible. The contextual variables used by Kinsella et al. are geographical coordinates extracted from geotagged "tweets" which allow them to model locations at varying levels of granularity. Using these contextual variables, prediction of the location of an individual "tweet" and of the location of a user is calculated.

3.6 Understanding Personal Interests Using Geographical Variables

The most known utilization of geographical variables in IR is multilingual text retrieval. In that context, the user's location is exploited for inferring the preferred language, style, and content from a multilingual collection of documents.

A more refined detection of the user's location permits a contextual search system to associate geographical variables to a specific user, thus

resolving the problem of describing the user — if the user is frequently located at the same place, whatever happens at that place is likely interesting for the user. This is the idea expressed by Bennett et al. [21], which differs from the idea expressed by Zhuang et al. [183] because the latter exploit geographical ontologies while Bennett et al. exploit the user's location. Bennett et al. first estimate the probability that a URL is located at a location represented by a longitude/latitude pair. The estimation data are obtained from data sets of which tuples associate the visited URLs and the location at which the page is visited. Then, some contextual variables about URLs and locations are computed. In particular,

$$\text{ENTROPY}(u) = \text{E}_g(-\log(p(g|u)))$$

and

$$\text{KL}(u, b) = \sum_g p(g|u) \log\left(\frac{p(g|u)}{p(g|b)}\right)$$

are computed where u is a URL, g is a location, and b is the background distribution. Location p can be a precise proxy of a user when, for example, it is provided by location sharing services, which allow users to voluntarily annotate the specific time that a user was at a location. When g is an effective proxy of user, these contextual variables can be used for detecting personal interests — see for example the work carried out by Cheng et al. [43].

Belkin et al. [18, 19] already illustrated that an information need derives from an ASK which in turn results from a problematic situation. When a problematic situation is a natural disaster, such as an earthquake or hurricane, the users' information needs become urgent and strong. When natural disasters are considered, considering the physical detachment of the user is indeed "natural" and becomes a crucial contextual variable affecting the personal interests — the farther the user is from the disaster, the less the user is interested in it.

As natural disasters heavily involve people, the user is likely to be more interested in such an event if he is connected with friends or relative involved by the event. Yom-Tov and Diaz [177] investigate how the users' information need is affected by their physical detachment,

which is estimated by their physical location in relation to that of the event, and by their social detachment, which is estimated by the number of their acquaintances who may be involved by the event.

Of the work carried out by Yom-Tov and Diaz we would like to emphasize the simple mathematical model which leverages language models. The elements of the model proposed by Yom-Tov and Diaz [177] are: topic t, aspect a, and user u. A topic is represented as a set of queries users may issue when seeking information about that topic, $Q(t)$, and the set of documents which may satisfy users seeking information about that topic, $R(t)$. An aspect is a part of a topic and therefore each topic may be partitioned into aspects. A user is interested in one or more aspects of a topic. The probability $p(a|u)$ that a user u is interested in aspect a can be estimated on the basis of different contextual variables. In particular, physical detachment and social detachment are contextual variables used to this end. Using these elements, Yom-Tov and Diaz [177] introduce the probabilities necessary to predict the user's personal interests. For example, let $U(d)$ be the set of users physically detached from an event by d, which is the geographic distance between the user and epicenter of the event, the probability that $u \in U(d)$ is interested in t can be estimated as

$$p(t|U(d)) = \frac{\sum_{u \in U(d)} |Q(u) \cap Q(t)|}{\sum_{u \in U(d)} |Q(u)|},$$

where $Q(u)$ is the set of us queries.

3.7 Concluding Remarks and Suggestions

Although many decades have passed since the early experimentation in query expansion and relevance feedback, query expansion and modification is still an effective method for personalizing the search engine result page to the user's personal interests. Therefore, before embarking on complex methods, query expansion is still worth investigation.

Nevertheless, the use of query expansion in the past research works, especially if implemented by relevance feedback, suffered from much

noise due to the addition of irrelevant features to, or the removal of meaningful features from the original query.

Within contextual search, this phenomenon can be termed lack of need of personalization, that is, little potential of personalization. Therefore, checking that personalization is worth implementation within a contextual search application is a helpful step.

Other sources of content, such as search engine result pages or documents, are a valuable sources for personal interest detection, thus, these contextual variables should be considered when designing methods for personal interest detection.

On the contrary, social variables are still not fully reliable since the social phenomena and the related research clashes with the issues of privacy and spam.

The diversification of a search engine result page may be an effective way to improving contextual search and adding personalization, however, the use of machine learning-based methods requires reliable, large data sets, some assumptions and a great deal of computational resources. This is especially true when interaction variables are utilized for detecting personal interests — dense data sets are needed for achieving reliable parameter estimation.

User modeling through profiles and categories may provide some additional power, however, at the expenses of generality and flexibility.

Implicit Relevance Feedback data may provide some additional input if these are carefully selected and checked; for example, display or dwell time may be a good proxy of personal interests, yet some data analysis is necessary to test whether this is case.

The paper written by Agichtein et al. [2] is rich in experimental results and useful suggestions. Other details on the Matthew effect are provided by Broder et al. [29]; as that paper addresses frequent queries, it should be read together with the paper written by Broder et al. [28] devoted to rare queries. Chakrabarti et al. [41] is worth reading because it focusses on a couple of issues: the lack of feedback on what happens when advertisements are displayed along with the search engine result pages but not followed by the users; the lack of content-based context when advertisement displays and click-through data are matched. Hence, the authors investigate Statistical Models

(statistical models) for integrating word-based page representations with Click-Through Data (click-through data) for predicting correct advertisement displays. The overall problem and approaches to tagging and linking large photocollections is, for example, described by Crandall and Snavely [46]. A well-written paper with some interesting ideas on surrogating the availability proprietary query logs is by Dang and Croft [50]. Diaz [53] clearly explains the methods and gives a background on probability that is rather helpful for understanding the content of many contextual search papers, although the topic is not easy to newcomers. Diaz [53] models as prior distributions of probability of the features and he makes useful connections: between distributed IR and news integration; to query classification (user intent detection); to the non-stationarity of the news vertical; topic detection; Information Filtering; active learning; mining bursts in query logs (burst detection); models of user click behavior. The thorough study reported by Di Buccio [52] addresses many theoretical, methodological, and experimental aspects of implicit relevance feedback — the literature survey complements this survey. A proof-of-concept tool for task context-based search that has been designed and evaluated is presented by Gyllstrom et al. [67]. The paper written by Hong et al. [73] is another research work relevant to the combination of social and geographical variables for personal interest detection. The paper by Joachims et al. [88] is one of the best ones on this subject as it contains useful information to be considered when designing user studies and analyzing click-through data. An investigation of the usefulness of display time is by Kelly and Belkin [96]. A knowledge discovery in databases-oriented paper is by Li et al. [116] where evaluation is based on simulation fueled by logs provided by a commercial search engine company. Although the employed algorithms may be computationally expensive, it is worth investigating if they can compete in a real setting and dynamic environment. The introduction and the technical sections are clearly written also for non-experts in machine learnings. Moreover, the pseudo-code algorithms are useful and necessary to understand and replicate the experiments. Liu et al. [121] report simple statistical methods. A detailed description of how to employ language model is reported by Luxenburger et al. [122]. Ma et al. [123] report a prototype

tool and the related evaluation. Melucci and Rehder [128] describe an extended Vector Space Model introduced for dealing with the specific problem of searching for information relevant to the researchers of a Space Agency, thus showing another example of contextual search in workplaces addressed in Section 2.4.3. The framework presented by Melucci and White [129] has further been generalized and reported by Melucci [125]. As for the Kendall τ, see for example Melucci [126]. The article by Pitkow et al. [137] is a good exposition of some points of this survey. An early innovative proposal of actual prototype is by Shen et al. [156]. The paper by Teevan et al. [160] is not a statistical paper in a strict sense, yet it addresses some basic problems of contextual search from an implicit relevance feedback point of view.

It is interesting to note that more complex statistical models do not seem to be superior to relevance feedback. The paper written by Teevan et al. [162] is a beautiful paper on evaluation. The technical part is confined to the methods and is useful for explaining the evaluation findings. The other paper written by Teevan et al. [161] belongs the same stream of research and gives results on whether the use of personal context is feasible or effective in a situation where the participant's behavior is affected by diverse hidden variables that, for example, reduce the willingness or the ability to provide additional information regarding their need. Teevan et al. [163] reports a heuristic method that worked for the authors' data set. Note that there are papers of theirs with similar algorithms. An interesting theoretical contribution is by Yue and Joachims [179].

4

Document Quality

Document: from Latin *docere*, "teach."
Quality: from Latin *qualis*, "of what kind."
Oxford Dictionary

4.1 Introduction

In noncontextual search, relevance is usually referred to a form of expression of the overall value of the document itself as a whole and no attention is paid to other qualities of a document than the relevance of the document to the user's need of information. This means that, for example, a non-contextual search system equally ranks both a document written by an effective expert of a topic and a document written by an amateur of the same topic. It would in contrast be of crucial importance to have some separate information on other qualities than relevance.

The fact that the characteristics of document quality other than relevance were not separately considered in the early days of IR can be explained by three facts: first, the collections that were managed by the IR systems at that time were always controlled; second, the user's information needs were somehow standardized and the IR systems were built by considering a restricted set of user requirements — any notion

of context was simply ignored; finally, quality and importance may vary depending on the contextual factor and these are becoming important in the modern contextual search systems.

In this section, we address document qualities other than relevance, in particular, we address authoritativeness, attractiveness, worthiness, and novelty. In the literature we have surveyed, these qualities have often been addressed using statistical models applied to contextual variables.

4.2 Detecting Document Quality Using Interaction Variables

4.2.1 Detecting News Novelty

In this section we describe some exemplar research works on detecting news novelty using interaction variables. Detecting news is somehow related to query intent or personal interest detection, since a contextual search system has not only to predict whether a query is about a specific document type, but also to predict whether this is the type of document wanted or of interest to the user. For example, there are queries expressing the need of news and a system should be able to respond with news: the intent is to have news while the documents should be qualified as news.

A common way followed by search engines is to respond with a classical search engine result page integrated with a news box on the top of the page. Once the news items are displayed, users decide whether to click on one of these news items. From a contextual search point of view, a system has to decide whether news items are worthy with respect to novelty; an old news item is much less useful than a fresh, novel news item.

Diaz [53] assumes that there is a strong correlation between the click-through rate of the news displayed along with a search engine result page and newsworthiness, such that the higher the number of users clicking on a news item, the higher the probability that the news item is novel. The other assumption is that similar queries relate to the same news. Diaz designs a classifier that, after appropriate training, can decide about news publishing with minimum error. The features

that are used at training time are the number of past queries very similar or identical to the current query and the number of documents that were viewed after both current and past queries were issued. By using these features, the parameters of a Bayesian statistical model are estimated; the adoption of the Bayesian statistics comes from the probabilistic nature of the parameters that drive the decision about news publishing.

According to Diaz, one potential drawback to the use of Bayesian statistics is maintaining a collection of dense language model vectors of previously seen queries. Even worse, given a new query, such a language model should be computed for each of these previously seen queries. In practice, he avoids much of this cost by only using the top terms and inverted indices for storing previously seen queries.

The problem of deciding whether news items are to be displayed becomes harder when events come one after the other at a speed that makes the reuse of past events useless or difficult. The problem of identifying breaking news is addressed by Li et al. [114] where the authors address the problem using a multi-armed bandit model, that is, a statistical model sequentially selecting news based on the information of the user and news, and adapting itself on the basis of the user's click-through data in the same way a player who selects one arm of a bandit out of the possible arms receives a payoff and sequentially learns how to select the next arm for maximizing the total payoff.

To detect news, the multi-armed bandit algorithm used proceeds in steps. At each step, the algorithm chooses an arm (i.e., a news item) and expects a payoff that depends on both the user and the item; the user decides whether to click the item and the system receives the actual payoff; the system learns the best item depending on the payoffs; the total payoff is defined as the sum of the payoffs received in the trial. The final goal of a multi-armed bandit algorithm is to maximize the total payoff.

4.2.2 Detecting Authoritativeness

In Section 4.4.2, we mentioned the PageRank algorithm proposed by Brin and Page [25] when illustrating the use of social variables for detecting page authoritativeness.

The WWW links are social variables exploited by PageRank as a source of evidence. Some investigations have been made in order to test whether interaction variables can substitute WWW links in detecting authoritativeness. Zhu and Mishne [181, 182] report a click-through data-based algorithm called ClickRank as a possible substitute of PageRank.

Suppose sessions have been extracted from a click-through data set. A session contains data about the user's behavior when interacting with a small set of WWW pages, performing a task, or querying with an intent.

The data that reveal the importance the user gives to the pages are, for example, dwell time on a page and the click order within a general trail of user activities: accessing one page before, more frequently, or for a longer time than another page in the same session may be interpreted as an endorsement by the user of the page in the session. Given these data, consider a session j and a page i. The session can be viewed as an ordered list of pages displayed in a search engine result page visited by the user. We have the ClickRank of i in j, that is,

$$\begin{aligned}&\text{CLICKRANK}(i,j)\\ &\quad= \sum_{\text{event } k \text{ of session } j} w_r(k,j)w_t(k,j)I(i \text{ is visited at event } k)\end{aligned}$$

where

$$w_r(k,j) = 2\frac{n_j + 1 - r(k)}{n_j(n_j+1)}$$

is the rank weight of event k of session j including n_j events, $r(k)$ is the rank of the page involved at k — the larger the $r(k)$ is, the lower the $w_r(k,j)$ is — and

$$w_t(k,j) = (1 - e^{-\lambda_d d(k)})e^{-\lambda_l l(k)}$$

is the temporal weight of event k of session j, $d(k)$ is the dwell time on the page at event k, $l(k)$ is the latency time of the page, and the λs are mean parameters. The larger $d(k)$ or the smaller $l(k)$ are, the larger the temporal weight function. I is the indicator function.

An example of the combined use of link analysis algorithms and interaction variables for detecting document authoritativeness is the

paper by Almeida and Almeida [6]. This paper presents a ranking function that combines two contexts: the content of the objects being retrieved and the community of the user. The authors exploit link analysis algorithms with user sessions; a user session is a sequence of accesses issued by a user during a single interaction with a service. A link between the sessions is simulated by a function matching the content of two nodes; the function is the cosine of the angle between the vectors that represent the nodes. The resulting graph is then described in matrix form and is processed through Hyperlinked Induced Topic Search (HITS) by Kleinberg [101]. Thus, each session is assigned two scores: the authority score and the hub score. The combination of two or more eigenvectors of the co-citation or the bibliographic coupling matrices produces the communities. Almeida and Almeida found that the top half of the search engine result page includes the best communities and the bottom half includes the worst communities.

4.3 Detecting Document Quality Using Content Variables

4.3.1 Detecting Attractiveness

By document attractiveness, we mean the quality of a document having beneficial features that induce to accept what is being offered through such a document without necessarily knowing the content of the document in advance.

The basic problem of detecting attractive documents is being able to automatically predict whether or not an individual document or an entire set of documents is attractive and not necessarily or only relevant enough to the user's information need to be added to a search engine result page.

The best known instance of document attractiveness regards contextual advertising. The problem is to automatically predict whether or not an individual ad or an entire set of ads is attractive enough to be displayed along with a search engine result page. In contextual advertising, attractiveness has an immediate and clear economical implication: the goal is to show a limited number of relevant ads and at the same time not drive the user away. It is often undesirable to

show many ads; on the contrary, it is sometimes desiderable to not show them.

We are concentrating on ad attractiveness in this section because it may correspond to a quality of a document delivered by a contextual search system which aims at picking out a few documents deemed particularly relevant, that is, attractive to the user; the correspondence is between ad and document and between ad attractiveness and document attractiveness to the user.

An ad ranking system is a specialized IR system selecting the best ads that match a query, are paid by customers and perhaps clicked by users. Suppose an ad ranking system provides a ranked ad page.

A simple approach to selecting attractive ads is to set a threshold and cut the bottom ranked ads off the list. A thresholding algorithm decides the top-ranked portion of the ad ranked page to be displayed. The lower the threshold is, the higher the level of coverage is, where coverage is defined as the proportion of queries for which at least one ad is displayed. Of course, different thresholds display different ads. If, for a given query, all of the ads have a very low score that is below the threshold, no ads are displayed. The primary advantage of thresholding is that it is very simple to implement, but the primary disadvantage is the need to choose a threshold. In fact, effectively modeling such tradeoffs appears to be a difficult problem.

Broder et al. [27] use a thresholding algorithm as the baseline for their machine learning-based method modeled as a support vector machine-based binary classifier. Given a query and the *set* of candidate ads, the goal is to predict whether or not the *entire set* of ads is relevant enough to be displayed. Therefore, the prediction mechanism takes a query and a set of ads as input and produces a yes/no decision as to whether the entire set should be displayed.

Clearly, the classification of entire sets of ads is more complex than a straightforward binary classification of individual ads and poses some challenges to evaluation. One reason is that training sets are usually available in a query/ad/judgment form and have to be processed so that a query/set of ads/judgment is available. A natural way to proceed is to average the individual judgment. However, an average judgment is a rational number and therefore a threshold is again necessary.

Broder et al. decided not to keep threshold aside as an external parameter, in contrast, they keep the threshold as an internal parameter of the model. The first step to building a classifier is feature selection. Broder et al. have selected the features for their support vector machine so that two factors are considered when a set of ads is classified. The first factor is the degree to which an ad is about a query (aboutness). These average features are computed under the assumption that these features will be useful for predicting whether an entire set of ads is about or not because the ad ranking system used them to decide whether an individual ad is about the query. The features that were considered are: number of words overlapping query and ad; cosine-based similarity between query and ad; probability of translation of words; the expected mutual information measure between a query word and an ad word; the χ^2-based measure of association between a query word and an ad word; the overall bid price of the set of ads. The other factor, which has been called cohesiveness, is the degree to which the ads of a set are mutually cohesive within the set. Broder et al. adopt two different yet related cohesiveness measures: clarity score and entropy, defined as follows, respectively:

$$\text{CLARITY}(p_a, p_c) = \sum_z p_a(z) \log \frac{p_a(z)}{p_c(z)}$$

and

$$\text{ENTROPY}(p_a) = \sum_z p_a(z) \log p_a(z),$$

where $p_a(z)$ is the probability of an ad z within a set of ads, and $p_c(z)$ is the probability of an ad z within entire collection of ads. Their paper reports the details of probability calculation.

The notion underlying CLARITY and ENTROPY has also been adopted by Zhou and Croft [180] for measuring the quality of a document as the divergence between the probability distribution over the words of a document and the probability distribution over the words of the "average" document of a collection. The rationale behind the measure proposed by Zhou and Croft, which we are calling QUALITY, is that low quality documents have unusual word probability distributions, that is, if a document differs significantly from the word usage

in an average document, the quality of this document may be low. In the QUALITY measure, the average document is represented by the collection language model. More concretely, the collection-document entropy of the word probability distribution in the collection is first computed as CLARITY(p_c, p_d), where p_c is a word probability distribution in the collection, $p_c(w)$ is the probability of an word w in the collection, $p(d)$ is the word probability distribution in the document d, and $p_d(w)$ is a probability of w in d. Moreover, the information-to-noise ratio INFO(p_d, p_c) for d and c is computed as as the total number of terms in the documents after indexing divided by the raw size of the document. Clarity (i.e., divergence) and information-to-noise ratio are "injected" into the probability that d is a high quality document in a collection c as follows:

$$\text{QUALITY}(p_d, p_c) = \text{CLARITY}(p_c, p_d)\text{INFO}(p_d, p_c).$$

Along the same line, Bendersky et al. [20] proposed a quality-biased ranking method that leverages features of document quality combined through a Markov random field.

A Markov random field is a probabilistic model such that a set of random variables are assigned a probability distribution which emphasizes strong dependencies between some small sets of variables while assuming independence between the other variables. A common view of a Markov random field is an undirected graph whose vertices are variables and edges are dependencies. The strong dependencies between some small sets of variables are cliques; an example is provided by Figure 4.1. Thus, the probability function of a Markov random field g is given can be as follows:

$$p(G) = \prod_{c \text{ is a clique of } G} \psi(c),$$

where G is the graph representing the Markov random field. The probability function $\psi(c)$ is usually written as an exponential family of

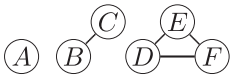

Fig. 4.1 Six random variables in a Markov random field. There are one, two, and three-node cliques.

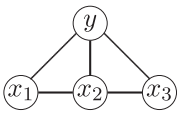

Fig. 4.2 A Markov random field with one document y and three query words x_1, x_2, x_3 where x_2 is adjacent to x_1 and x_3. Every word is tied with y.

probability functions:

$$\psi(c) = \exp \sum_{i=1}^{|c|} \lambda_i f_i(x_i),$$

where $|c|$ is the number of "cliqued" variables, f_i is the probability function of the ith variable x_i of clique c, and the λs are parameter.

A Markov random field can be used in IR as follows: given a document y_j and a query x, the document and the query words are modeled as the random variable vertices, and occurrence and adjacency are modeled as the edges. Figure 4.2 shows an example. The probability function can be estimated through the usual document-term frequencies and parameters. After applying the logarithms as usual, the scoring IR function for a document becomes

$$\sum_{c \text{ is a clique of } G} \sum_{i=1}^{|c|} \lambda_{i,j} f_{i,j}(x_i).$$

Suppose an arbitrary number of document quality features are selected and measured upon every document. The exponential family of Markov random field-based scoring function can be instantiated with additional retrieval functions and parameters that measure document quality. The quality-biased function becomes a sum of two functions: the usual content-based function and the quality-based function.

The variables of the quality-based function are the features about quality observed upon documents. The features used by Bendersky et al. [20] are: number of visible terms on the page (as rendered by a web browser) number of terms in the page title field; average length of visible terms on the page; fraction of anchor text on the page; fraction of visible text on the page (as rendered by a web browser); entropy of the page content; stopword/non-stopword ratio; fraction of terms in

the stopword list that appear on the page; the depth of the url path (number of backslashes in the url); fraction of table text on the page. In their paper, useful references are reported.

4.3.2 Detecting Worthiness

Efforts in understanding document worthiness using content variables have mainly been conducted through user studies, since human experts are able to assess document qualities and contribute to building training sets. An example of worthiness detection is the report written by Ng et al. [133] on an investigation of document worthiness within a newspaper and newswire domain. The qualities that are investigated in the paper are: accuracy; source reliability; objectivity; depth; author credibility; readability; verbosity/conciseness; grammatical correctness; one-sided/multiviews, and are viewed as complementary to relevance. Beside the classifiers that implement the decision about whether a document has a quality, Ng et al. provide information about the pattern with which the qualities are inter-related (see Table 4.1) and the most significant (p-value < 0.01) relationships between qualities and content variables (see Table 4.2). Three main document quality groups have been selected. The first group contains accuracy, source reliability, and credibility, and may be viewed as the document quality such that a proposition reported in the document is true. We term the first group "precision quality" (i.e., the quality of a

Table 4.1. Patterns of document quality reported by Ng et al. [133].

	Component	
Pattern	First	Second
Accuracy	0.689	0.206
Source reliability	0.604	0.158
Author credibility	0.615	−0.227
Objectivity	0.776	0.000
One-sided/multiviews	0.753	−0.106
Depth	0.787	−0.109
Readability	0.013	0.811
Verbosity/conciseness	−0.025	0.804
Grammatical correctness	0.073	0.729

Table 4.2. Correlation between document quality and content variables reported by Ng et al. [133]; r stands for the Pearson correlation coefficient.

Quality	Content Variable	r
Accuracy	Personal pronoun	−0.202
Source reliability	Distinct organization	0.154
Author credibility	Date unit, e.g., day, week	0.235
Objectivity	Possessive pronoun	−0.219
One-sided/multiviews	Past-form verb	0.238
Depth	Distinct organization	0.236
Readability	Closing parenthesis	−0.141
Verbosity/conciseness	Subordinating preposition or conjunction	−0.197
Grammatical correctness	Average length of paragraph in words	−0.172

document of being precise). The second group, which contains objectivity, one-sided/multiviews and depth, may be viewed as the document quality such that a true proposition has been reported in the document. We term the second group "recall quality" (i.e., the quality of a document of being exhaustive). The third group, which contains readability, verbosity/conciseness, and grammatical correctness, may be viewed as the document quality such that a proposition has been conveyed without errors. We term the term group "signal quality" (i.e., the quality of a document of being transmitted without error).

Each document quality has a strong predictor implemented by a content variable. The first group of document qualities (i.e., precision quality) can be predicted through the frequency of personal pronouns, distinct organization names, date units. The more distinct organization names or date units and the less personal pronouns there are, the higher the precision quality is. The second group of document qualities (i.e., recall quality) can be predicted through the frequency of possessive pronouns, past-form verbs, and distinct organization. The more past-form verbs or distinct organizations and the less possessive pronouns there are, the higher the recall quality is. Finally, the third group of document qualities (i.e., signal quality) can be predicted through the frequency of closing parentheses, subordinating prepositions or conjunctions, and the average length of paragraph in words. The fewer of these variables there are, the higher the signal quality is.

4.3.3 Detecting Document Readability

Readability can be defined as the quality of a document to be deciphered, that is, whether it is easy, quick or enjoyable to read. The specific quality depends on the user or on the type of user.

A significant example of user type is children. While research mainly concentrates on interface design and content ranking, the search engine result pages are prepared for an "average" user without taking care of the specific needs which depend on the user's age, cultural background, and education. While these issues for adult users can "easily" addressed, for example, when multi-lingual issues occur, these issues for children are less frequently and less easily addressed.

In general, readability can radically change the assessment of relevance of a document to the same extent authoritativeness did. However, the research work that has been done so far using content variables is relatively limited.

A contribution to detecting document readability is the paper written by Collins–Thompson et al. [45]. Although their focus is on search engine result page re-ranking, they also illustrate a methodology for detecting document readability and the user's ability to read a document. The user's ability to read a document is represented as an ordinal variable R_u that has as many values as the number of grades (the simplest case is binary such that a user either has or has not the ability to read a document). Document readability is similarly represented by an ordinal variable R_d. The larger the difference $(R_u - R_d)^2$, the less the readability of a document for a user. Collins–Thompson et al. assume that the probability that a user u likes a document d is proportional to $\exp\{(R_u - R_d)^2\}$. However, both R_d and R_u are unknown. To obtain an estimation, suppose R_d is known and a click-through data set is available. If $\exp\{(R_u - R_d)^2\}$ is associated to the click-through data such that the more d is clicked, the more likely u likes d, it is possible to estimate R_u. The document readability R_d can be estimated as the sum of the expected readabilities of the words occurring in d. The paper written by Collins–Thompson et al. [45] describe the estimation procedure.

4.3.4 Detecting News Novelty

The novelty track of TREC has been an early attempt to investigate models and methods for understanding the novelty of news using content variables. The task of the novelty track of TREC was as follows: Given a topic and a ranked search engine result page, an IR system had to find the relevant and novel sentences that should be returned to the user from this page. The task essentially required methods for passage retrieval, information filtering, and natural language processing. The basic approach was to select relevant sentences by measuring similarity to the topic whereas novel sentences were selected by dissimilarity to past sentences. In the research work reported in the proceedings of the novelty track of TREC, similarity and dissimilarity computation was mainly based on the vector space model or the BM25-based probabilistic model, Relevance Feedback was also an effective means, and, in some cases, part-of-speech analysis and named entity recognition were used with some success.

4.4 Detecting Document Quality Using Social Variables

4.4.1 Detecting News Novelty

In Section 4.3.4 we mentioned how standard IR methods can be adapted for detecting news novelty and that this adaptation has been reported by the novelty track of TREC. In Section 3.6, we mentioned how geographical variables can support detecting personal intents using the methods that have been developed within the situation caused by natural disasters, which can be viewed as a special case of new novelty. The Töhoku earthquake on March 11, 2011 is a glaring example of how natural disasters can be the ground where information is rapidly and extensively produced or consumed.

In this section, we describe how social variables can be exploited to spread news or gossip through neighbors or "friends" when, for example, natural disasters, riots or terrorist attacks cause an instantaneous need of information. Actually, this section does not specifically provide detection methods, yet it does provide some basic mathematical properties of the graphs from which the social variables are observed as well as some bibliographic references to further study this topic.

Social networks or micro-blogs are represented by using graphs where vertices are people and edges are connections between people. These graphs are called "random" because any non-trivial statement about them can be validated by only NP-hard algorithms, randomized algorithms or events of a probability space.

The graphs that represent social networks or micro-blogs share some well known properties reported, for instance, by Barabási and Albert [16] with the graph that represents the WWW.

The distribution of the frequency of the edges connected to the vertices follows a power-law such that there are k^{-3} vertices connected by k edges. These graphs also exhibits the so-called Matthew effect such that the rich get richer (see also Section 3.3), that is, a new vertex entering the graph connects to the most connected vertices. Moreover, the diameter of these graphs is relatively very low even when the size is very large, in particular, its order of magnitude is logarithmic in the number of vertices.

The diameter of the graph representing a social network is an upper bound to the number of steps (or rounds) necessary for transmitting information from one vertex to another vertex.

The crucial difference between social networks or micro-blogs and the WWW is that people know each other whereas pages do not. This feature reflects on the type and then on the mathematical properties of the graphs representing social networks, micro-blogs and the WWW.

The graphs that represent social networks or micro-blogs are instances of the preferential attachment graphs introduced by Barabási and Albert [16] such that the probability that an edge is followed depends on the mutual knowledge between the vertices.

In contrast, the graphs that represent the WWW are random graphs such that the probability that an edge is followed only depends on the number of edges of each vertex.

When information is spread across social networks or micro-blogs, the number of steps necessary to transmit information from one vertex to another vertex is much lower than the diameter. In particular, Doerr et al. [54] found that a constant c exists such that the number of steps necessary to transmit information from one vertex to another vertex is bounded by $c \log(n)/\log(\log(n))$. Doerr et al. explain that, given

two vertices exchanging information, a third vertex with less edges than and connected to the other two vertices may exist such that information runs more rapidly between the two nodes through the third node than directly between the two nodes. This phenomenon is due to the preferential attachment policy followed by these graphs: the third node has less connecting edges than the other two nodes and therefore requires less time to spread the information, the time being proportional to the number of connecting edge.

4.4.2 Detecting Authoritativeness

The large size of the WWW has caused the storage of pages which should be classified at different grades of importance in the search engine indexes and then in the search engine result pages. In general, not only the relevance of a document, but also its importance is crucial to the user's need of information. In a search engine result page, a document written by an expert of a topic should be ranked by an IR system higher than a document written by an amateur of the same topic. In this section, we describe how social variables can be exploited to detect authoritative or popular WWW pages. Most of the section is devoted to link analysis-based methods since WWW can be viewed as the most known social contextual variable.

At first approximation, it is necessary to separate "good" relevant documents from "bad" yet still relevant documents (the question whether non-relevant documents should be classified according importance is left apart).

Many researchers have explored the idea that the degree to which a document is considered important by the user depends on the degree to which the document has been considered important by other users. This is the idea that makes the utilization of social variables relevant to this monograph and in particular to this section.

There are basically three main types of social variables exploited to detect importance: the paths leading to WWW pages implemented by the statically stored WWW links; the answers given to questions by more or less expert users; the "tweets" forwarded to connected "friends."

The number of paths leading to a WWW page implemented by the statically stored WWW links through the pages authored by other users can estimate the importance of the page. This consideration has led to the design and implementation of link analysis algorithms that represent the WWW as a graph and estimate the importance of the content of a document on the basis of the number of links which point to it directly or indirectly.

Most link analysis algorithms that represent the WWW as a graph are rooted in the fact that many users of WWW search engines want both authoritative and relevant documents. However, even if the IR systems were able to retrieve many relevant documents it would not be automatic that these documents were authoritative.

The links between WWW pages can become a source of evidence for distinguishing authoritative documents from the others. The use of the number of paths to a page to measure the authority of the page is based on the idea that, if the author of page i thinks that page j is an authority, he is likely to insert a link from i to j. The presumption is that the reverse of the implication mentioned above is valid, that is, a link is likely to express the authority of the target page.

The PageRank algorithm proposed by Brin and Page [25] is the most glaring example of the link analysis algorithm. PageRank basically measures the degree to which a document is reachable by all the other documents — the more the page is reachable, the more it is authoritative.

The PageRank metaphor is quite well known. Suppose a user randomly surfs the WWW, where "randomly" means that the user chooses the next link with probability depending on the number of out-going links and on a damping factor d. With probability $1 - d$, the user chooses one link out of the links out-going from the current page; the probability is based on the number of out-going links. With probability d, he chooses any other page. The damping factor controls the trade-off between the user's will to follow the links and the alternative will to exit from the paths stored in the WWW. It can be shown that the user reaches the most authoritative pages and that the probability that he still reaches them if he carries on surfing is stable after a sufficiently large number of steps. The number of links that point to a document

is a necessary measure of authority in PageRank, however, it is not a sufficient measure: a popular document, for example, a commercial search engine homepage is directly pointed to by many links without being an authority, and authoritative documents might conversely be less pointed-to than popular documents.

What makes an authority different from a popular document is that the documents which point to authorities are likely to be authorities as well, whereas the ones which point to popular documents are likely not to be. Therefore, a model trying to capture authority has to consider that the greater the authority of a document is, the greater the authority of j is. PageRank can measure this relationship since: the more the documents with links pointing to j (in-links of j) there are, the greater the authority of j is; the greater the authority of the documents with links pointing to j is, the greater the authority of j is.

Whereas PageRank can only detect authoritative documents, in contrast, Hyperlinked Induced Topic Search (HITS), which has been proposed by Kleinberg [101], can detect another type of document, that is, hubs. A hub document has many links to authoritative pages. HITS detects hubs on the basis of the following argument: the more authoritative documents with in-links from j (out-links of j) there are, the more j is a hub; the more hub documents with out-links to i (in-links of i) there are, the more i is authoritative.

The WWW links are not necessary to confer authority to a document. One might infer authority on the basis of "stand-alone" properties of WWW documents, for example, typographical features or layout: a well designed and written document is likely to be considered more authoritative than one written confusedly. In general, document genre is an effective source of evidence for contextual search.

Although the effectiveness of the search engines which implements PageRank might still depend on factors other than the algorithm itself, it is quite well accepted that PageRank can partially be an effective means for making IR effective.

The idea of using links to measure authority is not really new: bibliometrics is based on the calculation of the number of citations among papers in order to assess the importance of a journal. The main difference is that the notion of the WWW page is radically different from

that of a journal paper: whereas bibliographic citations refer to whole documents, links refer to documents that might be part of a larger document. Finally, a link might not express the authority of the target document, an authoritative document might not be linked to by any link or a link might point to one or more subjects. What characterizes PageRank from HITS is the independence of the queries issued by the end users. This means that PageRank is precomputed at indexing time before any query is submitted to the search engine.

There have been some proposals to enhance PageRank in order to incorporate statistical model to represent and to adapt the algorithm to the user's context. The versions of these proposals have been termed "personalized" or "topic-sensitive." In this regard, Haveliwala [71] addresses the problem due to the presence of heavily linked pages which get highly ranked for queries for which they have no particular authority. These pages are considered important in some subject domains, yet they may not be considered important in others, regardless of what keywords may appear either in the page or in anchor text referring to the page.

When crawling the WWW, the topic-sensitive PageRank algorithm generates as many vectors as the predefined topics. Each vector is different from the next due to the different damping factors. The computational cost is minimized because vector calculation is done offline. At retrieval-time, topic-sensitive PageRank takes the linear combination of the topic-sensitive vectors, weighted using the similarities of the query (and any available context) to the topics. The main finding reported by Haveliwala is that the average precision for the rankings induced by PageRank biased toward topics that are supposed to be interesting to a user is higher than that of the unbiased PageRank.

Two results should be considered as a warning directed to anyone aiming to use PageRank or HITS for ranking documents by authoritativeness. The first result by Melucci and Pretto [127] refers to PageRank. They found that the actual computation of PageRank performed through a power series on a large WWW directed graph tends to hinder variations in the order of large rankings, presenting a high stability in its induced order both in the face of large variations of the damping factor value and in the face of truncations in its computation. The

second result by Peserico and Pretto [136] refers to HITS. They found that the algorithm can converge slowly, but not too slowly, both in score and in rank. Moreover, they found that an exponential number of iterations might be necessary and sufficient to converge to a ranking (and a score) that is even remotely accurate.

The answers given to questions by more or less expert users is a different approach to detecting document quality. The difference of this approach does not lie in the contextual variables, on the contrary, it lies in the information objects indexed and searched by its implementation, that is, the Aardvark search engine which was proposed by Horowitz and Kamvar [74]. Instead of searching documents, expert users are searched by that social search engine — an expert user is most likely able to answer the question asked by another user. The mathematical model is simple and inspired by language models. Similar to the model by Yom-Tov and Diaz [177], the elements of the model proposed by Horowitz and Kamvar [74] are: answer t, question q, and user u. The probability that a user u is the expert user answering question q is

$$p(u|q) = \sum_{t \in T} p(u|t)p(t|q),$$

where T is the set of possible answers. Then, the probability that a user u can successfully answer user vs question is $p(u|v)$. The latter is a measure of the authoritativeness of u with respect to v. If u is an author of a document, us authoritativeness can be transferred to the document and the model proposed by Horowitz and Kamvar [74] can be used for detecting document quality.

The "tweets" forwarded to connected "friends" are at the same time a source of noise and evidence to assess the degree to which "tweets" are accurate, true and reliable. "Tweets" are special documents which particularly suffer from lack of credibility since they can quickly be produced in such large quantities that a manual filter cannot efficiently and effectively work.

The lack of credibility of "tweets" is the theme addressed by Castillo et al. [39] who investigated whether meaningful features can be extracted from these social media for detecting unreliable "tweets." Castillo et al. propose a set of features that refer to the user's

ability in assessing credibility. In particular, they assume users express sentiments according to their own opinions and background when they receive a "tweet." Moreover, users are assumed to be capable of assessing whether a "tweet" is worth propagation to other users. Finally, users are assumed to be capable of assessing the authoritativeness of the cited URLs, if any. Castillo et al. also define a set of quantitative features such as the fraction of tweets that contain URLs, the fraction of tweets with hashtags, the fraction of sentiment positive and negative in a set, registration age, number of followers, number of followees, and the number of the user's "tweets." The experimental results reported in their paper show that the classical textual features are very effective in automatically classifying "tweets" by credibility. Examples are the URLs occurring in the "tweets." Another outcome is that poorly credible "tweets" are posted by users who frequently posted "tweets" in the past. Finally, credible "tweets" are associated with frequent propagation — the more the "tweets" are propagated, the more credible these "tweets" are. The results reported by Castillo et al. suggest that the social network of users act as social filter; a user can exploit his personal judgment for stopping unwanted "tweets."

4.5 Concluding Remarks and Suggestions

The most accessed literature on contextual search has less considered document quality a contextual factor than query intent (or search task) and personal interests (or personalization) which are on the contrary viewed the main pillars of contextual search.

In contrast, we think document quality has both a great potential in determining relevance thanks to a mature suite of statistical models to detect different aspects (e.g., authoritativeness, attractiveness, worthiness, and novelty) of document quality.

A reason that makes document quality a contextual factor worth investigating is the variety of statistical models and the related literature designed to detect quality, for example, link analysis algorithms and part-of-speech analysis algorithms.

The other and perhaps most crucial reason is the centrality of document in determining the user's relevance assessment constrained by

context: a document is the container of information relevant to the user's information need and the shape providing this information may varyingly fit the contextual factors. In other words, we think that the role played by document quality in contextual search is a exemplar situation when form is more important than content.

The paper written by Brody and Kantor [31] is an example of the user study-based approaches to understanding document worthiness. Further improvements in detecting authoritativeness using social networks can be achieved if the proposal reported by Cataldi et al. [40] to weight users by PageRank is adopted. One of the most relevant papers on bibliometrics is by Garfield [65]. In the context of social tagging, some improvement is reported by Harvey et al. [69] where it is stated that the topic modeling approaches provides better resource rankings than even the most competitive baselines. Kim et al. [99] describe another approach to detecting readability. Another HITS-like algorithm called SALSA is described by Lempel and Moran [111]. Random graphs were introduced by Erdös and Rényi [58, 59] and are explained in Palmer [135]'s tutorial book. The documentation about the novelty track of TREC can be found at the National Institute of Standards and Technology [132] (http://trec.nist.gov).

5

Contextual Search Evaluation

Evaluation: from Latin *ex-* "out, from" and
valere "to value, to be worth."
Oxford Dictionary

5.1 Introduction

In this section, evaluation is addressed by referring to the approaches published in the literature surveyed in this survey. However, producing a definitive survey of evaluation techniques is out of the scope of the monograph because excellent surveys have already been published on this topic. Thus, we have chosen to focus on actual experimented techniques, with the evaluation material in a supporting role rather than as the principal focus.

The literature of this survey is already rich with experiments and plenty of results are available to reach a preliminary opinion of the current contextual search methods.

Looking at the literature, the variety of approaches to evaluation in contextual search is striking — these approaches are more varied than those in traditional IR. This may be due to the lack of a well recognized evaluation methodology analogous to the Cranfield methodology by

now accepted in the IR community, which can be applied to the variety of contexts or the notions of context addressed in the papers.

5.2 Evaluating Contextual Search Using Content Variables

The main content variables are document content and query content. From an evaluation point of view, it should not be a surprise that these contextual factors correspond to document and query which are the pillars of the Cranfield methodology adopted in IR evaluation. However, document, query, and assessment should in contextual search be taken in quite a broad sense because there is not yet a well accepted evaluation methodology for contextual search as the Cranfield methodology is for classical IR. Thus, although document refers to multimedia data containers in IR, we extend this concept to log records or "live" WWW pages. Moreover, query is not only meant as a representation of a well-formed user's information need but also as an implicit expression of this need — in this sense, a click may be viewed as a query.

In most research papers on contextual search, the evaluation is implemented using document sets. A document set is a corpus of written texts, spoken texts (i.e., speech), audio, images, video, music on a particular subject. What features a corpus is curation — data stored in a corpus are selected, organized and looked after over time by curators. Not every document set is a corpus, for example a source from which documents are extracted without being stored in a corpus is the "live" WWW, which cannot be considered a corpus in a strict sense because there is not any curation.

In contextual search, the traditional test collections are of little use because their design was independent of any context; this is not negative in itself, on the contrary, it was an explicit requirement for making these collections useable in different experimental contexts. However, this requirement hinders contextual search evaluation when methods for contextual factor detection need experimentation. We mention two aspects in this section.

The dualism between public availability of and proprietary right to data sets is a significant aspect of document sets in IR and even more so in contextual search evaluation. It is well known that there are

document sets that are not publicly available regardless of whether they are free of charge. The lack of public availability of document sets might hamper the reproducibility of the results since one can reproduce the methodology by which the results are obtained while producing other results with another data set. On the other hand, stopping research only because the data will not be made available to the public might limit the advance of Science. This dualism is significant in IR evaluation, yet it comes up even more significant in contextual search evaluation because the criteria used to select, organize, and look after data are influenced by context. Therefore, a researcher who employs these data for testing hypotheses on context has to know exactly how and what influences the data for the purposes of separating what is actually under experimentation and what is not.

Usually, corpora are publicly available whereas logs are not. The main reason for this is because of the commercial origin and the legal restriction to invading user's privacy of logs whereas corpora are often excerpts of already published newswire agencies, newspapers or broadcasting companies and are curated by non-profit organizations that aim at making knowledge available to the research community — note that publicly available does not imply free of charge.

Another aspect of document sets is the method used for building them. There two broad classes of methods, that is, automatic and non-automatic. Automatic methods are performed by software tools while non-automatic methods envisage some manual intervention. For example, logs are generated by automatic methods whereas test document collections are often created after being built by means of crawlers. Clearly, automatic methods make the production of large sets easier than manual methods, yet automatically produced sets require a more time consuming curation than manual sets with which curation and production often go together. The situation is varied. For instance, the TIPSTER collections and the WWW track of TREC collections are produced in a semi-automatic way — these collections are first compiled through some crawling, then, errors are fixed and the documents are assigned an identifier.

The literature in contextual search reports on evaluation with genuine queries, with recycled queries and with surrogate queries.

Genuine queries are only prepared for the particular purpose of an experiment and therefore are the most genuine kind of query with respect to the aims of the specific contextual search experimental setting. Genuine queries are issued by the participants to user studies in which they were asked to browse and search and then either directly or indirectly issue queries, for example by: manually selecting ten hashtags based on popularity; browsing fifty topics TREC test collections and picking five or seven most interesting topics; choosing a query to mimic a search the participants had performed earlier that day in order to closely mirror the searches that the participants conducted in the real world; selecting a query from a list formulated of general interest queries, then describing their intent and finally rating the relevance of documents relative to their intent — these queries had to be of general interest in order to provide the participants with results that would have some meaning for them.

Recycled queries are queries utilized in other experiments, therefore, they are less realistic because are produced in contexts different from the context of which search is evaluated. Recycled queries have been: generated from TREC topics; randomly extracted, selected from the most frequent queries or from those issued in a temporal interval, and stored in search engine logs.

Surrogate queries are data that were not intended as representation of an information need when they produced, yet they were used for this purpose in the experiment. Surrogate queries have been: extracted from all sessions in the user browsing behavior data; compiled from the set of tags of each test set of bookmarks; synthetically generated; selected from taxonomies.

5.3 Evaluating Contextual Search Using Interaction Variables

Context is the basic difference between contextual search evaluation and non-contextual search evaluation. The making of a judgment about the effectiveness of an IR system is conditioned by a contextual factor in the former case, whereas it is not in the latter case. As a consequence, the relevance judgments are made by experts only when building a

test collection for non-contextual search, whereas the relevance judgments should be made whenever context changes. Obtaining relevance judgments is impractical, since the costs and turnaround times are excessively high, and it is ill-suited to context-sensitive IR, since the information needs are not labeled by contextual factors.

Whereas corpora are the main collection of data for obtaining relevance judgments, in contextual search logs are used as frequently as corpora in various forms such as click-through data, queries, interaction data, user behavior data. Logs are regular or systematic collections of records of events, for example, requests, responses, incidents, observations, failures. Logs have been used in various fields of Computer Science for keeping memory of past interactions between, for example, users and systems and for predicting future events before they actually happen and cause damages. Besides some anonymization, logs are not curated, yet they are noticeable due to their large size mainly because they are automatically produced by servers of millions of clients. Logs play an important role in contextual search evaluation, since they often provide evidence of implicit assessment of relevance and other contextual factor. Therefore, logs can store the implicit feedback directly from the users.

The implicit feedback obtained directly from the users can be used to replace the relevance judgments obtained by experts and it may "capture" the contextual factor affecting the users when judging document relevance. Chapelle et al. [42] leverage this assumption and propose replacing the traditional Cranfield approach to collecting relevance judgments with an approach based on the collection of the implicit feedback obtained directly from the users. However, this is not Chapelle et al.'s main contribution, in our opinion. Rather, they propose interleaving a search engine result page produced by a system with the results of a search engine result page produced by another system, before presenting the results to the user. In this way, the user is unaware of the search engine result page from which a result comes and the degree of preference of a search engine result page to the other can be measured.

Although interleaving is easily implemented, it cannot provide an absolute measure of effectiveness, since a search engine result page

preferred to another can still be a poor retrieval result. Nevertheless, interleaving is repeated for the queries of an experimental data set, the researcher is provided with one (a_i, b_i) for each query i such that if $a_i > b_i$, the researcher infers a preference of system a, if $a_i < b_i$, the researcher infers a preference of system b, otherwise, he infers no preference and declare a tie. If A is the number of queries when a is preferred to b, B is the number of queries when b is preferred to a and T_{AB} is the number of ties, the following statistic can be computed:

$$\Delta_{AB} = \frac{A - B}{2(A + B + T_{AB})},$$

where the denominator is the number of queries. System a preferred to b when $\Delta_{AB} > 0$. Suppose three systems A, B, C are evaluated with the same number of queries. If $\Delta_{AB} > 0$ and $\Delta_{BC} > 0$ then $\Delta_{AC} > 0$. Moreover, $A - B$ is a frequency distribution of the possible number of queries such that a is preferred to b in a given number of queries in each of which there is the same probability of success, there is, it is a Binomial random variable. It follows that the hypothesis that a is better than b can be tested.

Whether interaction variables are good predictors of contextual factors and in general of relevance is still debated. An answer to this question is summarized in Table 5.1 whereas a more detailed answer is given in this section. A striking feature of the interaction variables is user behavior. No other component playing in a contextual search system has such a feature. Although emotions are still difficult to detect and further research is still needed for making emotions useable in contextual search, some user behavior can be measured and effectively incorporated within the search process. Hence, a key question is whether interaction variables can effectively be used in a large-scale, contextual search system in comparison with traditional content-based ranking functions or more advanced yet well experimented methods such as anchor text or link analysis algorithms.

Most of the research work on the use of click-through data for predicting personal interests and personalizing search engine result page lies on the assumption that click-through data are useful predictors of personal interest and then at least partially of relevance. This

Table 5.1. A summary of the arguments that are reviewed in Section 5.3.

Are interaction variables good predictors of personal interest?	
Yes	No
A striking feature of the interaction variables is user behavior and user behavior is intrinsically composing context.	A set of interaction variables does not necessarily refer to an individual user and may refer to a group or individual users who change over time behind a proxy.
Interaction variables appear to hold potential for capturing the differences between users when assessing relevance.	User experimentation has shown that a single interaction feature can vary greatly between users who tend to click the most clicked or the top-ranked results.
Interaction variables provide the feedback a user is reluctant to explicitly provide.	Explicit relevance judgments are more reliable especially those about the bottom ranks.
Interaction variables collection is unobtrusive.	Interaction variables anyway refer to humans, thus raising privacy issues and limiting the diffusion of publicly available data sets and the scientific comparison.
The click-through data improves retrieval effectiveness.	The improvement provided by click-through data is not at all consistent across all queries, large quantities of data are necessary, and the integration with other data types is useful.
The click-through data sets are inexpensive to collect.	More than half of the test queries had no click-through data and very frequent adult queries are unuseable.
The click-through data sets are large.	The click-through data sets are noisy.
The click-through data detect differences between users when interacting with the top-ranks of a search engine result pages.	Explicit assessments seem more likely to successfully predict relevance than click-through data because the latter are often unavailable for the bottom ranks while explicit assessments are.

assumption (and hope) derives from some empirical observations that the assessments that are explicitly given by the users significantly vary between the users who issued the same query, thus suggesting that the need for personalization should be detected at browsing time when the user clicks on the search engine result page — the users do not tell what they wish from the system when issuing the query, they tell it

when "talking" to the system. It is indeed when the users "talk" to the system that their behavior signals some differences in their intent behind the query.

As a general result interaction variables appear to be useful to hold potential for capturing the differences between users when assessing relevance, while classical content-based data and measures (e.g., similarity, probability of relevance of a document to a query) appear to hold potential for capturing variation across individuals, according to Teevan et al. [161].

The other reason that make interaction variables frequently utilized in contextual search is that interaction variables are often necessary. The vast literature about relevance feedback is pervaded by a common belief, that is, when explicit relevance feedback is proposed to a user, he is reluctant to explicitly provide feedback information independently of the form used to provide them.

There are diverse reasons for this reluctance: the user is called on to provide additional data while the benefits are not always obvious; the user is worried about privacy issues because the final use of the additional data appears opaque; the explicit provision of additional data always requires some additional design and implementation activities; interaction variables in principle implies the provision of data while the user is interacting with the system, thus bypassing the user's reluctance; the additional data can be given as numbers and not as qualitative contextual variables such as demographic data, thus reducing the risk of misuse and the issues of privacy; the user would not be disappointed if the benefits had not lived up to his expectations since he was not promised anything.

Although it may seem surprising, the click-through data improves retrieval effectiveness according to Chakrabarti et al. [41] who report that the click-through data significantly improves the accuracy of sponsored search results compared to traditional relevance scoring models that are solely based on semantic (e.g., cosine) similarity when the fraction of displayed sponsored results that are clicked is up to 0.2.

Beside effectiveness, the click-through data sets are also inexpensive to collect and some experimental research works showed that a careful design and detailed implementation allow to collect many useful data

about user interaction at no cost and preserving the user's privacy. Finally, the lack of difference between assessments with *ad-hoc* queries and assessments with recycled queries mentioned above in this case turns out to be an advantage, because the explicit assessments of the bottom of search engine result pages returned in response to recycled queries are as reliable as the explicit assessments of the bottom of search engine result pages returned in response to *ad-hoc* queries. This comparability between performances allows the recycled queries to be reused many times within different experiments.

There are, however, a number of drawbacks in using interaction variables for detecting personal interest and in general relevance. The pseudo relevance assessments based on click-through data have to be carefully considered before viewing them as a reliable proxy of relevance in the unlikely event of clicking results ranked in the bottom half of the search engine result page. If such attention were not paid, a shortage and sparsity of data would occur and estimation would be little reliable.

It is true that human users generate interaction variables, however, interaction variables may automatically be generated[1] or a set of interaction variables does not necessarily refer to an individual user and may refer to a group or individual users who change over time behind a proxy. However, even if the user would be more precisely identified, it has been shown through user experimentation that a single contextual variable could vary greatly between users and search tasks according to Kelly and Belkin et al. [96].

Moreover, although interaction variables appear to hold potential for capturing variation across individuals, the click-through data are more frequent for the top of search engine result page than would be expected given the explicit assessments, while the bottom of search engine result pages receive significantly fewer clicks than the explicit assessments, according to Teevan et al. [161].

The evidence that click-through data seems rather unreliable due to the even distribution and the consequent sparsity of data of the items ranked low in the search engine result page — yet they are far from useless — is confirmed by a study by Joachims et al. [88] who report

[1] We are not addressing spamming in this survey.

that users tend to click more frequently on the items in the top half of search engine result page than on the items in the bottom half of search engine result page, thus making click-through data insufficient for estimation and prediction and putting the question what would happen to reranking if there were additional click-through data on the items in the bottom half of search engine result page.

In this respect, explicit assessments seem more likely to successfully predict relevance than pseudo assessments like click-through data because a training set may contain enough assessments made on the bottom of search engine result pages that the users might not provide if they had to only click on those results. The same Joachims et al. [88]'s study report that the click-through rate significantly drops after the tenth item and decreases around the sixth item due to the need to go to the next search engine result page or vertically scroll the display, respectively. Interestingly, the drop is not uniform — it is higher within the top half than the bottom half of the search engine result page. After reversing the search engine result pages order, the items that are placed in the bottom half of the original search engine result page are significantly scanned more frequently than the top half search engine result pages items. Joachims et al. analyze how users interact with search engine result pages, how their behavior can be used as a proxy of relevance, and how eye-tracking data can help understand how users behave on search engine result pages in order to generate feedback from clicks. To evaluate the degree to which feedback signals indicate relevance, they compare the implicit feedback data against explicit feedback data they collected manually. The difference between the pseudo assessment provided by click-through data and explicit assessment is not consistent across different research works, for example, Joachims et al. [88, 89] found that the feedback generated from clicks shows reasonable agreement with the explicit assessments of the WWW pages.

The click-through data does not always improve retrieval effectiveness according to Agichtein et al. [2] who report that incorporating additional interaction variables in WWW search more markedly improves over the state-of-the-art (i.e., BM25) than using click-through data alone as a proxy of relevance, yet they also found that the improvement over the state-of-the-art is not at all consistent across all queries

and that more than half of the test queries had no click-through data. The results shows that individual clicking decisions are not only influenced by the relevance of the items, but that users behavior also depends on the decision made by the system about the ranking — this is crucial in contextual advertising and paints a bleak picture of the user's ability of a good and autonomous judgment. Nevertheless, Attenberg et al. [11] showed that the situation is a bit more complex than it seems at first sight especially within a contextual advertising domain.

Although a key advantage of click-through data sets is that these are collected in much larger quantities, the inexpensive means of collection of click-through data might not always be an advantage. Good quality can be achieved with a small quantity of data and manual control whereas it is a large quantity that might cause a low level of quality control, thus leading to noisy data. Nevertheless, the collection of data from user behavior may be controlled by appropriate statistical techniques that reduce the amount of noise.

In relation to the possibility of inexpensively collecting large click-through datasets, another sensitive yet related issue of the use of click-through data is the diffusion of proprietary data set and the lack of publicly available collections. As a consequence, many studies of query reformulation based on query logs nearly all make use of proprietary query logs and click-through data. Clearly, this lack makes comparison barely feasible and causes delays in their use within contextual search.

Another challenge issued by click-through data is that in a realistic IR system these data are more noisy than the data collected within a smaller and controlled environment. To address the issues of noise in click-through data, Dang and Croft [50] suggest using anchor text to simulate the queries and, as a proof of the concept, they construct a simulated query log from the anchor text in the TREC .gov2 test collection. A paper that explores whether interaction can be helpful in realistic environments has been written by Agichtein et al. [2] who exploit statistical models for WWW search and IR. As explicit human relevance judgments are available for a set of WWW search queries and results, the authors use a supervised machine learning technique to learn a ranking function that best predicts relevance judgments.

Sponsored results tend to have lower click-through rates than the organic results displayed in a search engine result page as showed by Attenberg et al. thanks to the detailed data collected by a navigation plug-in. This outcome is in line with those previously mentioned — anything that is not displayed on the top-left of the search engine result page is condemned to a low click-through rate and the user is sensitive to what he perceives as "organic." Once clicked, the WWW sites linked in sponsored search engine result page are more variably visited and more quickly abandoned than the WWW linked in organic Search Engine Result Pages (search engine result pages).

Moreover, the amount of activity varies across the topics both for trail length and duration. More interestingly, click-through data collected after a topic has been searched by the user is not necessarily associated with display time.

Click-through is on the contrary associated with query frequency — the most frequent queries resulting in many more clicks than infrequent queries — thus suggesting that the most frequent queries are those giving the most effective search engine result page and therefore the most active click-through.

The correlation between the level of click-based activity and the click-through rate is almost absent according to Attenberg et al.'s study. This means that click-through rate and "intensity" of click-through activity cannot be a signal of the same personal interest. This appears to be true for both sponsored and organic results. This result implies that while predicting personal interests using click-through rate may attract the users who may be really interested, a higher click-through rate does not translate to additional search activity once the user "landed" on the linked page. This outcome may be explained by the hypothesis that there are sponsored results which are more attractive than other sponsored results since once users click sponsored results they leave them almost immediately, according to Attenberg et al. [11].

The difference in click-through rate can also be measured by the larger variance of the number of clicks taken by users who access sponsored results than the trail lengths taken by users who access search engine result pages. A significant drop in the average expected

click-through rate at the early stage has also been observed by Li et al. [114], whose authors point out that like most feedback-based approaches, click-through data are usually more reliable with large amounts of historical data and less so for the ones with little or no history. Similarly, click-through data effectiveness can be improved if these data are combined with high click-through rate queries given that historic and low click-through rate queries are detrimental to the accuracy.

5.4 Concluding Remarks and Suggestions

We conclude this section suggesting some additional reading. Agosti [3] reports some guidelines on evaluation within DLs which is a natural area where contextual search may be applied. Almeida and Almeida [6] use a company intranet repository. Anastácio et al. [8] address semi-anonymity when geographical variables are exploited. Bai et al. [14, 15] use TREC collections. Bian et al. [23] use LETOR and TREC collections. Broder et al. [27] use corpora. Campbell et al. [34] use a company intranet repository. Campbell et al. [34] use corpora produced from company intranets. Cao et al. [35] use the ACM KDD Cup data set. Chapelle et al. [42]'s research work report interleaving as an alternative approach to collecting relevance assessments, since the conventional Cranfield-based approach to evaluation is not free of drawbacks although it is the most used and well accepted in non-contextual search. Dai et al. [49] use "live" WWW. Dang and Croft [50] use TREC collections. Dang and Croft [50]'s work is an example of careful design and detailed implementation allow to collect many useful data about user interaction at no cost and preserving the user's privacy. Diaz [53] uses "live" WWW. Finkelstein et al. [61] use a corpus extracted from a CD-ROM. Freund et al. [62] use corpora produced from company intranets. Guo and Agichtein [66]; Harvey et al. [69] use "live" WWW. Haveliwala [71] uses a corpus extracted from a WWW site. Hawking and Craswell [72] report on using the WWW track of TREC `.gov` collection. Hu et al. [75] use logs (around 20 million WWW queries collected from around 650,000 Web users). Hu et al. [76] use a corpus extracted from a WWW site. Ingwersen [77] reports on evaluation

from both an information seeking and retrieval and operational point of view. Jansen and Spink [85] use logs (nine major commercial search engine anonymized and well prepared query logs); see also Jansen and Spink [85]; Jansen et al. [82, 83, 84]. Jansen [81]'s paper is a useful side-effect is the public availability of the data set. Joachims [86] uses corpora. Kelly [94] provides a complete account on some approaches to interactive IR evaluation illustrated by Ingwersen. Lau et al. [108, 109] use TREC collections. Li et al. [117, 118] illustrate an interesting approach to automatic training set construction. Liu et al. [121] use a series of small data sets that have been built with user cooperation. Ma et al. [123] use "live" WWW. Sanderson [150] surveys the most general issues of Cranfield style-based evaluation. Shen et al. [156] use TREC collections. Spink [159] discusses the potential of user behavior and interaction variables. Teevan et al. [160] use "live" WWW. Yue and Joachims [178] use TREC collections. van Rijsbergen [167] and the publications cited in Section 1.4 are worth reading from an evaluation point of view.

6

Conclusions

It is our opinion that modeling and implementing context is necessary in IR, yet this is not sufficient and is sometimes counterproductive for improving the performances of a system. Modeling and implementing context is necessary because one cannot *a priori* say that the evidence observed in addition to the conventional queries cannot be exploited for improving the performances of a system. However, the use of the sources of evidence for contextual search should be decided from time to time. This decision should be weighed up with the costs and the benefits because, for example evidence is often noisy. The need for this decision implies that modeling and implementing context is not a sufficient means for improving IR effectiveness, since additional research areas such as Economics, Cognition, Physics should be explored to help the IR researchers to better understand contextual search.

We think that additional definitions or speculations of context are no longer necessary. The literature devoted to this topic is by now large enough. A pragmatic perspective of contextual search like the one used in this monograph that helps researchers to implement context in IR would be preferable. In this monograph, we are suggesting to adopt two notions, that is, contextual variable and contextual factor

as the constituent parts of a statistical view of context which condenses both the numerical and the conceptual aspects of context into a single notion. The statistical models are numerous and if they are not complex the computational power is abundant for most tasks. The systematization of contextual search according to contextual factors, contextual variables and statistical models has not been straightforward, since there is no consensus of opinion among researchers and their papers betray a cacophony of systematizations of contextual search. Nevertheless, the systematization given in this monograph is quite satisfactory because reflects what happens in Science: contextual variables are observed and statistical models estimate or predict unobservable contextual factors.

We think that the statistical models are necessary in contextual search because they model data in large quantities for the purpose of inferring patterns in a context from the data in a representative sample. Although such models are not sufficient for providing a complete and precise picture of the context under examination, the experimental findings suggest that these models can be improved by additional sources of evidence.

The experimental findings suggest that the simplest statistical models are more likely effective and are certainly more efficient than complex models. Investing in sophisticated algorithms is often interesting from a theoretical and computational point of view, yet it sometimes turns out to be little effective from an IR point of view — however, these investigations are always worthwhile from a purely scientific perspective and should always be encouraged because this often yields important (and unexpected) scientific discoveries.

It is our opinion that the combination of statistical models has a great scientific potential to increase the effectiveness of contextual search systems. For example, the combination of, say, a linear model with a Markov chain might turn out to be more effective than a complex non-linear model. Research is focussed on the investigation and comparison of one or two statistical models, but the additional predictive power that might be produced when these models are linked together is still an open question.

Textual content-based IR systems are relatively easy to design since text is simply scanned and processed while contextual data are simply ignored. A challenge of contextual search is that what to observe and not only how to process the contextual data must be determined and therefore treated by an IR system. This challenge is similar to that encountered in non-textual content-based IR and in particular in image or video indexing and retrieval where robust and efficient computer vision indexing methods are not yet available and textual descriptions (e.g., tags, annotations) are still the main means of content description.

We think that the investigation of non-textual sources of evidence and content is a great opportunity for the researchers in contextual search. For example, user behavior data is a source of evidence which is not necessarily expressed in and then is not constrained by the ambiguity of the natural language. A new approach to IR would open up.

Acknowledgments

I would like to thank the Information Management Systems research group, led by Maristella Agosti, at the Department of Information Engineering for the continuous collaboration; Fabrizio Sebastiani for inviting me to write this survey; Doug Oard for his great patience and encouragement; and three anonymous reviewers for their careful and thoughtful comments.

This work has partially been funded by the EU 7th Framework Programme Marie Curie IRSES project N. 247590 "QONTEXT."

A

Implementations

A.1 SearchPad

SearchPad is an extension to the search engine result pages described by Bharat [22]. SearchPad maintains contextual variables and in particular containing interactions variables short-term user's queries and hyperlinks of search engine result pages the user likes. Other implicit relevance feedback data are: the order in which result pages are viewed; display time; page visits. In this way relationships between queries and favorite links are maintained and can be cross-referred to the implicit relevance feedback data. SearchPad allows experienced users to: search on many unrelated topics in parallel and with many browser windows; keep track of different sessions; refine queries; post a query to different search services; manage relevance assessments. SearchPad parallels classical bookmarking since the latter cannot: store temporary interesting links with long-term favourite links; allow the user to easily repost queries; effectively manage multiple search sessions. As is customary, SearchPad is independent of the client operating system and most processing is done on the client side.

A.2 IntelliZap

The starting point of IntelliZap is that query intents can be represented by the context of the query words surrounding them — the underlying assumption is that some text around the query words must exist, the latter being not valid for WWW pages which might not contain any text. IntelliZap is a client–server, meta-search Information Retrieval (IR) system described in Finkelstein et al. [61] — "client" and "server" are used in the paper yet the system runs on the user's computer. The client-side of the system reads a word highlighted by the user and extracts the textual window around the word — this window is called context in the paper and is an example of content-based contextual variable in this survey. The server-side classifies the contextual variable into predefined categories; it sends the contextual variable to WWW search engines according to the selected categories; lastly, it combines and reranks the Search Engine Result Pages (search engine result pages) received by the WWW search engines. To perform reranking, the authors use a semantic network and a related metric that returns a score reflecting the degree to which the meanings of two contextual variable are related.

The experiments designed by the authors aimed at measuring the difference between IntelliZap and four major search engines in terms of number of relevant results in a search engine result page. Twenty-two subjects were presented with some topics each and had to search the WWW by using IntelliZap and four major search engines. IntelliZap outperformed all the other search engines with one-word queries. In another experiment, twelve subjects were presented with five topics. Each subject was assigned a random search engine. IntelliZap achieves a level of performance only comparable to major search engines and its response time is longer than that of the conventional search engines due to meta-search. Actually, recall would have been a more appropriate measure than precision in the event of meta-search, yet recall can hardly be measured in a WWW-based setting.

A.3 GroupBar

GroupBar is a desktop-resident toolbar which allows users to arrange windows into groups and to switch between tasks with a single mouse

click Smith et al. [157]. A user of GroupBar adds tiles to the group and can arrange with a meaningful order or a correspondence between the position of a tile and the position of the represented window on the screen. This correspondence stems from the authors' observation that the larger the display surfaces, the more the users leave more applications running and grouped in the associated windows. A key feature of GroupBar is that it allows users to perform operations on all of a group simultaneously.

The authors report in Smith et al. [157] on a field study utilized for evaluating the usefulness and usability of GroupBar in comparison with the existing TaskBar. The five participants played roles developed to represent target user groups. The authors chosen for *in situ* method to establish how important the new GroupBar features were for the participants. If, after approximately one week of use, participants were using GroupBar, this would provide evidence that GroupBar is superior to TaskBar. The authors also report on a user study involving eighteen participants. During the evaluation, the experimenter interrupted the participants to switch between tasks at given moments. The authors measured task time, subjective satisfaction responses to a post-search questionnaire, and overall tool preference. During the evaluation, the experimenter interrupted the participants to switch between tasks at given moments. The field study gave results above the average but no comparison or statistical significance testing were performed. The user study gave borderline results since a one-tailed t-test on task time revealed a task advantage at $p = 0.07$. The post-search questionnaire and overall satisfaction are markedly in favor of GroupBar.

A.4 Stuff I've Seen

Stuff I've Seen (SIS) is a system designed for information re-use Dumais et al. [56]. Perhaps, the key aspect of SIS is the uniqueness of a central index containing the complete description of every information object of a collection. In this way, a user has a complete collection with objects of any type uniformly described. The other aspect that is related to the preceding aspect is that contextual search is performed on this collection and then all the information objects are uniformly

treated when SIS searches for, matches against a query and ranks the information objects relevant to the user's information need. Other tool (e.g., IntelliZap or Y!Q described in Sections A.2 and A.5, respectively) accomplish this task by, for example, accessing various WWW search engines and re-ranking objects on the client side.

The evaluation of SIS involved some hundred people who used the system for a few weeks. A search log revealed that the users behaved as they do with WWW search engines (i.e., short queries, almost total absence of Boolean operator, resource, places or people finding queries). A post-search questionnaire also revealed that the highest scores given to SIS by the users were those typically given to effective WWW search engines (i.e., search tools are essential, search engine result page previews are useful, finding things is easy) while the lowest scores referred to the need of advanced search functions. In the middle of the ranking of what the users liked are the functions related to desktop search (i.e., e-mail message search, intranet search, exploring the WWW by using SIS).

A.5 Y!Q

The authors of Kraft et al. [104] present Y!Q, that is, a search application integrated with a major WWW search engine. The observation behind Y!Q is that queries are formulated while the user is engaged in some task and that the users that issue these queries have some intents. What Y!Q aims to do is to dynamically capture the intents behind a query. Like IntelliZap, Y!Q uses a semantic network for modifying the query by using contexts, which are also in this case text windows, and sends the rewritten query to various search engines. Lastly, Y!Q also computes what the authors call contextual ranking.

Y!Q was the tool used in the research reported in Kraft et al. [103]. Following the idea that the user formulates a query when engaged in a task, contextual search is, according to these authors, a proactive process that suggests terms to the user. To this end, the authors experiment three algorithms to perform automatic contextual search. Query Rewriting (QR) is query expansion where terms are taken from content-based contextual variable. Rank Biasing (RB) generates

a content-based contextual variable by using some heuristics. Iterative Filtering Metasearch (IFM) generates multiple subqueries which are independently processed before the search engine result page are merged. The experiments that have been carried out for assessing Y!Q performance established that not only precision can be increased but also recall can be markedly increased without losing precision by using IFM. Indeed, IFM is based on meta-search, which is also utilized in IntelliZap, and therefore it is intended to increase recall.

A.6 PCAT

In Ma et al. [123], Personalized Categorization System (PCAT) is illustrated. The authors propose a search engine result page result categorization system which is tailored to personal user professional interests and skills. The underlying argument is that a traditional search engine result page does not present the results in the best way for every user and a combination of categorization and personalization would be more effective than a plain list of results. It is interesting to note how search engine result page diversification has been later studied for addressing this problem. From a methodological point of view, PCAT is based on the ODP category taxonomy. Although comparing search engine result pages with lists of categories in which documents are in turn ranked is a challenging task, the authors report on some experiments in the paper which show that PCAT can outperform a plain list or a non-personalized categorization system for many different tasks where effectiveness is measured as the average rank at which relevant documents are found.

References

[1] S. Abrol and L. Khan, "Twinner: Understanding news queries with geo-content using Twitter," in *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR '10)*, New York, NY, USA, pp. 10:1–10:8, 2010. ISBN 978-1-60558-826-1. doi: 10.1145/1722080.1722093. URL http://doi.acm.org/10.1145/1722080.1722093.

[2] E. Agichtein, E. Brill, and S. Dumais, "Improving Web search ranking by incorporating user behavior information," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '06)*, New York, NY, USA, pp. 19–26, 2006. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148177. URL http://doi.acm.org/10.1145/1148170.1148177.

[3] M. Agosti, "Digital libraries," in *Advanced Topics in Information Retrieval*, (M. Melucci and R. Baeza-Yates, eds.), Information Retrieval. Springer, 2011.

[4] M. Agosti, R. Colotti, and G. Gradenigo, "A two-level hypertext retrieval model for legal data," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '91)*, New York, NY, USA, pp. 316–325, 1991. ISBN 0-89791-448-1. doi: http://doi.acm.org/10.1145/122860.122892.

[5] M. Agosti, N. Ferro, E. Panizzi, and R. Trinchese, "Annotation as a support to user interaction for content enhancement in digital libraries," in *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '06)*, New York, NY, USA, pp. 151–154, 2006. ISBN 1-59593-353-0. doi: http://doi.acm.org/10.1145/1133265.1133296. URL http://doi.acm.org/10.1145/1133265.1133296.

[6] R. B. Almeida and V. A. F. Almeida, "A community-aware search engine," in *Proceedings of the International Conference on World Wide Web (WWW '04)*, New York, NY, USA, pp. 413–421, 2004. ISBN 1-58113-844-X. doi: http://doi.acm.org/10.1145/988672.988728.

[7] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2004.

[8] I. Anastácio, B. Martins, and P. Calado, "Using the geographic scopes of Web documents for contextual advertising," in *Proceedings of the Workshop on Geographic Information Retrieval (GIR '10)*, New York, NY, USA, pp. 18:1–18:8, 2010. ISBN 978-1-60558-826-1. doi: 10.1145/1722080.1722103. URL http://doi.acm.org/10.1145/1722080.1722103.

[9] A. Ashkan and C. L. Clarke, "Characterizing commercial intent," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '09)*, New York, NY, USA, pp. 67–76, 2009. ISBN 978-1-60558-512-3. URL http://doi.acm.org/10.1145/1645953.1645965.

[10] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo, "Classifying and characterizing query intent," in *Proceedings of the European Conference on IR Research on Advances in Information Retrieval (ECIR '09)*, Berlin, Heidelberg, pp. 578–586, 2009. ISBN 978-3-642-00957-0. URL http://dx.doi.org/10.1007/978-3-642-00958-7_53.

[11] J. Attenberg, S. Pandey, and T. Suel, "Modeling and predicting user behavior in sponsored search," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '09)*, New York, NY, USA, pp. 1067–1076, 2009. ISBN 978-1-60558-495-9. doi: http://doi.acm.org/10.1145/1557019.1557135.

[12] L. Azzopardi, "Incorporating context within the language modeling approach for ad hoc information retrieval," PhD thesis, University of Paisley, 2005.

[13] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in *Proceedings of the International Conference on World Wide Web (WWW '08)*, New York, NY, USA, pp. 357–366, 2008. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367546. URL http://doi.acm.org/10.1145/1367497.1367546.

[14] J. Bai, J.-Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '07)*, New York, NY, USA, pp. 15–22, 2007. ISBN 978-1-59593-597-7. doi: http://doi.acm.org/10.1145/1277741.1277747.

[15] J. Bai, D. Song, P. B. andJ. Y. Nie, and G. Cao, "Query expansion using term relationships in language models for information retrieval," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '05)*, New York, NY, USA, pp. 688–695, 2005. ISBN 1-59593-140-6. doi: http://doi.acm.org/10.1145/1099554.1099725.

[16] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999. doi: 10.1126/science.286.5439.509. URL http://www.sciencemag.org/content/286/5439/509.abstract.

[17] D. Bartholomew, F. Steele, and I. Moustaki, *Analysis of Multivariate Social Science Data*. Statistics in the social and behavioral sciences series. CRC Press, 2008. ISBN 9781584889601.

[18] N. J. Belkin, R. Oddy, and H. M. Brooks, "ASK for information retrieval. Part 1: Background and theory," *Journal of Documentation*, vol. 38, pp. 61–71, 1982. ISSN 0022-0418.

[19] N. J. Belkin, R. Oddy, and H. M. Brooks, "ASK for information retrieval. Part 2: Results of a design study," *Journal of Documentation*, vol. 38, pp. 145–164, 1982. ISSN 0022-0418.

[20] M. Bendersky, W. B. Croft, and Y. Diao, "Quality-biased ranking of Web documents," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '11)*, New York, NY, USA, pp. 95–104, 2011. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935849. URL http://doi.acm.org/10.1145/1935826.1935849.

[21] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz, "Inferring and using location metadata to personalize Web search," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '11)*, New York, NY, USA, pp. 135–144, 2011. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2009938. URL http://doi.acm.org/10.1145/2009916.2009938.

[22] K. Bharat, "Searchpad: Explicit capture of search context to support Web search," *Computer Networks*, vol. 33, pp. 493–501, June 2000. ISSN 1389-1286. doi: http://dx.doi.org/10.1016/S1389-1286(00)00047-5. URL http://dx.doi.org/10.1016/S1389-1286(00)00047-5.

[23] J. Bian, T.-Y. Liu, T. Qin, and H. Zha, "Ranking with query-dependent loss for Web search," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '10)*, New York, NY, USA, pp. 141–150, 2010. ISBN 978-1-60558-889-6. doi: http://doi.acm.org/10.1145/1718487.1718506.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning and Research*, vol. 3, pp. 993–1022, 2003. ISSN 1532-4435. doi: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[25] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks ISDN Systems*, vol. 30, pp. 107–117, April 1998. ISSN 0169-7552. doi: http://dx.doi.org/10.1016/S0169-7552(98)00110-X. URL http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

[26] A. Broder, "A taxonomy of Web search," *SIGIR Forum*, vol. 36, pp. 3–10, September 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792552. URL http://doi.acm.org/10.1145/792550.792552.

[27] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras, "To swing or not to swing: learning when (not) to advertise," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '08)*, New York, NY, USA, pp. 1003–1012, 2008. ISBN 978-1-59593-991-3. doi: http://doi.acm.org/10.1145/1458082.1458216.

[28] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang, "Robust classification of rare queries using Web knowledge," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '07)*, New York, NY, USA, pp. 231–238, 2007.

ISBN 978-1-59593-597-7. doi: http://doi.acm.org/10.1145/1277741.1277783. URL http://doi.acm.org/10.1145/1277741.1277783.

[29] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," in *Proceedings of the International Conference on World Wide Web (WWW '02)*, Amsterdam, The Netherlands, May 2002. URL http://www9.org/.

[30] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel, "Search advertising using Web relevance feedback," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '08)*, New York, NY, USA, pp. 1013–1022, 2008. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458217. URL http://doi.acm.org/10.1145/1458082.1458217.

[31] S. Brody and P. Kantor, "Automatic assessment of coverage quality in intelligence reports," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers — vol. 2, (HLT '11)*, Stroudsburg, PA, USA, pp. 491–495, 2011. ISBN 978-1-932432-88-6. URL http://dl.acm.org/citation.cfm?id=2002736.2002834.

[32] C. Buckley and E. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR '04)*, pp. 25–32, 2004.

[33] G. Cai, "Relevance ranking in geographical information retrieval," *SIGSPATIAL Special*, vol. 3, pp. 33–36, July 2011. ISSN 1946-7729. doi: 10.1145/2047296.2047304. URL http://doi.acm.org/10.1145/2047296.2047304.

[34] D. R. Campbell, S. J. Culley, C. A. McMahon, and F. Sellini, "An approach for the capture of context-dependent document relationships extracted from bayesian analysis of users' interactions with information," *Information Retrieval*, vol. 10, pp. 115–141, 2007. ISSN 1386-4564. doi: http://dx.doi.org/10.1007/s10791-006-9016-2.

[35] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, and Q. Yang, "Context-aware query classification," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '09)*, New York, NY, USA, pp. 3–10, 2009. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1571945.

[36] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-aware query suggestion by mining click-through and session data," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '08)*, New York, NY, USA, pp. 875–883, 2008. ISBN 978-1-60558-193-4. doi: http://doi.acm.org/10.1145/1401890.1401995.

[37] N. Cardoso and M. J. Silva, "Query expansion through geographical feature types," in *Proceedings of the 4th ACM workshop on Geographical information retrieval (GIR '07)*, New York, NY, USA, pp. 55–60, 2007. ISBN 978-1-59593-828-2. doi: 10.1145/1316948.1316963. URL http://doi.acm.org/10.1145/1316948.1316963.

[38] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys*, vol. 44, pp. 1:1–1:50, January 2012. ISSN 0360-0300. doi: 10.1145/2071389.2071390. URL http://doi.acm.org/10.1145/2071389.2071390.

[39] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of the International Conference on World Wide Web (WWW '11)*, New York, NY, USA, pp. 675–684, 2011. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963500. URL http://doi.acm.org/10.1145/1963405.1963500.

[40] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proceedings of the International Workshop on Multimedia Data Mining (MDMKDD '10)*, New York, NY, USA, pp. 4:1–4:10, 2010. ISBN 978-1-4503-0220-3. doi: 10.1145/1814245.1814249. URL http://doi.acm.org/10.1145/1814245.1814249.

[41] D. Chakrabarti, D. Agarwal, and V. Josifovski, "Contextual advertising by combining relevance with click feedback," in *Proceedings of the International Conference on World Wide Web (WWW '08)*, New York, NY, USA, pp. 417–426, 2008. ISBN 978-1-60558-085-2. doi: http://doi.acm.org/10.1145/1367497.1367554.

[42] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue, "Large-scale validation and analysis of interleaved search evaluation," *ACM Transactions on Information Systems*, vol. 30, pp. 6:1–6:41, March 2012. ISSN 1046-8188. doi: 10.1145/2094072.2094078. URL http://doi.acm.org/10.1145/2094072.2094078.

[43] Z. Cheng, J. Caverlee, K. Y. Kamath, and K. Lee, "Toward traffic-driven location-based Web search," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '11)*, New York, NY, USA, pp. 805–814, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063693. URL http://doi.acm.org/10.1145/2063576.2063693.

[44] Z. Cheng, B. Gao, and T.-Y. Liu, "Actively predicting diverse search intent from user browsing behaviors," in *Proceedings of the International Conference on World Wide Web (WWW '10)*, New York, NY, USA, pp. 221–230, 2010. ISBN 978-1-60558-799-8. doi: http://doi.acm.org/10.1145/1772690.1772714.

[45] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, "Personalizing Web search results by reading level," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '11)*, New York, NY, USA, pp. 403–412, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063639. URL http://doi.acm.org/10.1145/2063576.2063639.

[46] D. Crandall and N. Snavely, "Modeling people and places with internet photo collections," *Queue*, vol. 10, pp. 30:30–30:44, May 2012. ISSN 1542-7730. doi: 10.1145/2208917.2212756. URL http://doi.acm.org/10.1145/2208917.2212756.

[47] W. Croft and J. Lafferty, eds., *Language Modeling for Information Retrieval,* volume 13 of *Kluwer International Series on Information Retrieval*. Kluwer Academic Publishers, 2002.

[48] W. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

[49] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li, "Detecting online commercial intention (OCI)," in *Proceedings of the International Conference on World Wide Web (WWW '06)*, New York, NY, USA, pp. 829–837, 2006.

ISBN 1-59593-323-9. doi: http://doi.acm.org/10.1145/1135777.1135902. URL http://doi.acm.org/10.1145/1135777.1135902.

[50] V. Dang and B. W. Croft, "Query reformulation using anchor text," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '10)*, New York, NY, USA, pp. 41–50, 2010. ISBN 978-1-60558-889-6. doi: http://doi.acm.org/10.1145/1718487.1718493. URL http://doi.acm.org/10.1145/1718487.1718493.

[51] G. De Francisci Morales, A. Gionis, and C. Lucchese, "From chatter to headlines: harnessing the real-time Web for personalized news recommendation," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '12)*, New York, NY, USA, pp. 153–162, 2012. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124315. URL http://doi.acm.org/10.1145/2124295.2124315.

[52] E. Di Buccio, "Design, implementation and evaluation of a methodology for utilizing sources of evidence in relevance feedback," PhD thesis, University of Padua, 2011.

[53] F. Diaz, "Integration of news content into Web results," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '09)*, New York, NY, USA, pp. 182–191, 2009. ISBN 978-1-60558-390-7. doi: http://doi.acm.org/10.1145/1498759.1498825.

[54] B. Doerr, M. Fouz, and T. Friedrich, "Social networks spread rumors in sublogarithmic time," in *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC '11)*, New York, NY, USA, pp. 21–30, 2011. ISBN 978-1-4503-0691-1. doi: 10.1145/1993636.1993640. URL http://doi.acm.org/10.1145/1993636.1993640.

[55] D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Understanding the relationship between searchers' queries and information goals," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '08)*, New York, NY, USA, pp. 449–458, 2008. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458143. URL http://doi.acm.org/10.1145/1458082.1458143.

[56] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've seen: A system for personal information retrieval and re-use," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '03)*, New York, NY, USA, pp. 72–79, 2003. ISBN 1-58113-646-3. doi: http://doi.acm.org/10.1145/860435.860451. URL http://doi.acm.org/10.1145/860435.860451.

[57] E. Efthimiadis, "Query expansion," in *Annual Review of Information Science and Technology (ARIST),* vol. 31, chap. 4, (M. Williams, ed.), Medford, NJ, pp. 121–185, 1996.

[58] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, 1959.

[59] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences (A Mathematikai Kutató Intézet Közleményei)*, vol. 5, pp. 17–61, 1960.

[60] W. Feller, *An Introduction to Probability Theory and its Applications,* volume 1. Wiley, 3rd ed., 1968.

[61] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: the concept revisited," *ACM Transactions on Information Systems*, vol. 20, pp. 116–131, 2002. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/503104.503110.

[62] L. Freund, E. G. Toms, and C. L. Clarke, "Modeling task-genre relationships for IR in the workplace," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '05)*, New York, NY, USA, pp. 441–448, 2005. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076110.

[63] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, "Analysis of geographic queries in a search engine log," in *Proceedings of the International Workshop on Location and the Web (LOCWEB '08)*, New York, NY, USA, pp. 49–56, 2008. ISBN 978-1-60558-160-6. doi: 10.1145/1367798.1367806. URL http://doi.acm.org/10.1145/1367798.1367806.

[64] V. Ganti, A. C. König, and X. Li, "Precomputing search features for fast and accurate query classification," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '10)*, New York, NY, USA, pp. 61–70, 2010. ISBN 978-1-60558-889-6. doi: http://doi.acm.org/10.1145/1718487.1718496. URL http://doi.acm.org/10.1145/1718487.1718496.

[65] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, vol. 178, pp. 471–479, 1972.

[66] Q. Guo and E. Agichtein, "Towards predicting Web searcher gaze position from mouse movements," in *Proceedings of the International Conference Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, pp. 3601–3606, 2010. ISBN 978-1-60558-930-5. doi: http://doi.acm.org/10.1145/1753846.1754025. URL http://doi.acm.org/10.1145/1753846.1754025.

[67] K. Gyllstrom, C. Soules, and A. Veitch, "Activity put in context: Identifying implicit task context within the user's document interaction," in *Proceedings of the International Symposium on Information Interaction in Context (IIiX '08)*, New York, NY, USA, pp. 51–56, 2008. ISBN 978-1-60558-310-5. doi: http://doi.acm.org/10.1145/1414694.1414707.

[68] P. Halmos, *Finite-dimensional Vector Spaces, Undergraduate Texts in Mathematics*. New York, USA: Springer, 1987.

[69] M. Harvey, I. Ruthven, and M. J. Carman, "Improving social bookmark search using personalised latent variable language models," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '11)*, New York, NY, USA, pp. 485–494, 2011. ISBN 978-1-4503-0493-1. doi: http://doi.acm.org/10.1145/1935826.1935898. URL http://doi.acm.org/10.1145/1935826.1935898.

[70] A. Hassan, R. Jones, and F. Diaz, "A case study of using geographic cues to predict query news intent," in *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*, New York, NY, USA, pp. 33–41, 2009. ISBN 978-1-60558-649-6. doi: 10.1145/1653771.1653780. URL http://doi.acm.org/10.1145/1653771.1653780.

[71] T. H. Haveliwala, "Topic-sensitive PageRank," in *Proceedings of the International Conference on World Wide Web (WWW '02)*, New York, NY, USA, pp. 517–526, 2002. ISBN 1-58113-449-5. doi: http://doi.acm.org/10.1145/511446.511513.

[72] D. Hawking and N. Craswell, "The very large collection and Web tracks," in *TREC: Experiment and Evaluation in Information Retrieval,* chap. 9, (E. Voorhees and D. Harman, eds.), MIT Press, 2005.

[73] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the Twitter stream," in *Proceedings of the International Conference on World Wide Web (WWW '12)*, New York, NY, USA, pp. 769–778, 2012. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187940. URL http://doi.acm.org/10.1145/2187836.2187940.

[74] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *Proceedings of the International Conference on World Wide Web (WWW '10)*, New York, NY, USA, pp. 431–440, 2010. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772735. URL http://doi.acm.org/10.1145/1772690.1772735.

[75] D. H. Hu, Q. Yang, and Y. Li, "An algorithm for analyzing personalized online commercial intention," in *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD '08)*, New York, NY, USA, pp. 27–36, 2008. ISBN 978-1-60558-277-1. doi: http://doi.acm.org/10.1145/1517472.1517476.

[76] J. Hu, G. Wang, F. Lochovsky, J.-T. Sun, and Z. Chen, "Understanding user's query intent with Wikipedia," in *Proceedings of the International Conference on World Wide Web (WWW '09)*, New York, NY, USA, pp. 471–480, 2009. ISBN 978-1-60558-487-4. doi: http://doi.acm.org/10.1145/1526709.1526773. URL http://doi.acm.org/10.1145/1526709.1526773.

[77] P. Ingwersen, "The user in interactive information retrieval evaluation," in *Advanced Topics in Information Retrieval*, (M. Melucci and R. Baeza-Yates, eds.), Springer, 2011.

[78] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

[79] W. H. Inmon, *Building the Data Warehouse*. Wiley, 1996.

[80] W. H. Inmon, "The data warehouse and data mining," *Communications on ACM*, vol. 39, pp. 49–50, November 1996. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/240455.240470. URL http://doi.acm.org/10.1145/240455.240470.

[81] B. J. Jansen, *Instructions for Obtaining Search Engine Transaction Logs*. 2011. URL http://faculty.ist.psu.edu/jjansen/academic/transaction_logs.html.

[82] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the user intent of Web search engine queries," in *Proceedings of the International Conference on World Wide Web (WWW '07)*, New York, NY, USA, pp. 1149–1150, 2007. ISBN 978-1-59593-654-7. doi: http://doi.acm.org/10.1145/1242572.1242739. URL http://doi.acm.org/10.1145/1242572.1242739.

[83] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of Web queries," *Information Processing and Management*, vol. 44, pp. 1251–1266, May 2008. ISSN 0306-4573.

doi: 10.1016/j.ipm.2007.07.015. URL http://portal.acm.org/citation.cfm?id=1351187.1351372.

[84] B. J. Jansen, D. L. Booth, and A. Spink, "Patterns of query reformulation during Web searching," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 1358–1371, July 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:7. URL http://portal.acm.org/citation.cfm?id=1568763.1568772.

[85] B. J. Jansen and A. Spink, "How are we searching the world wide web?: A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, pp. 248–263, January 2006. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2004.10.007. URL http://dx.doi.org/10.1016/j.ipm.2004.10.007.

[86] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '02)*, New York, NY, USA, pp. 133–142, 2002. ISBN 1-58113-567-X. doi: http://doi.acm.org/10.1145/775047.775067.

[87] T. Joachims, *The Support Vector Machine Software Package*. http://svmlight.joachims.org, 2012.

[88] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '05)*, New York, NY, USA, pp. 154–161, 2005. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076063. URL http://doi.acm.org/10.1145/1076034.1076063.

[89] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search," *ACM Transactions on Information and Systems*, vol. 25, April 2007. ISSN 1046-8188. doi: 10.1145/1229179.1229181. URL http://doi.acm.org/10.1145/1229179.1229181.

[90] C. B. Jones and R. S. Purves, "Geographical information retrieval," *International Journal of Geographical Information Science*, vol. 22, pp. 219–228, 2008. doi: 10.1080/13658810701626343. URL http://www.tandfonline.com/doi/abs/10.1080/13658810701626343.

[91] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating query substitutions," in *Proceedings of the International Conference on World Wide Web (WWW '06)*, New York, NY, USA, pp. 387–396, 2006. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135835. URL http://doi.acm.org/10.1145/1135777.1135835.

[92] D. Kelly, "Measuring online information seeking context. Part 1: Background and method," *Journal of the American Society in Information Science and Technology*, vol. 57, no. 13, pp. 1729–1739, 2006. doi: http://dx.doi.org/10.1002/asi.v57:13.

[93] D. Kelly, "Measuring online information seeking context. Part 2: Findings and discussion," *Journal of the American Society in Information Science and Technology*, vol. 57, no. 13, pp. 1862–1874, 2006. doi: http://dx.doi.org/10.1002/asi.v57:14.

[94] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 1–224, 2009.

[95] D. Kelly and N. J. Belkin, "Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback during interactive information retrieval scrolling and interaction: exploring implicit sources of user preferences for relevance feedback," in *Proceedings of the Annual International ACM Conference on Research and Development in Information retrieval (SIGIR '01)*, New York, NY, USA, pp. 408–409, 2001. ISBN 1-58113-331-6. doi: http://doi.acm.org/10.1145/383952.384045. URL http://doi.acm.org/10.1145/383952.384045.

[96] D. Kelly and N. J. Belkin, "Display time as implicit feedback: Understanding task effects," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '04)*, New York, NY, USA, pp. 377–384, 2004. ISBN 1-58113-881-4. doi: http://doi.acm.org/10.1145/1008992.1009057.

[97] D. Kelly and X. Fu, "Elicitation of term relevance feedback: an investigation of term source and context," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '06)*, New York, NY, USA, pp. 453–460, 2006. ISBN 1-59593-369-7. doi: http://doi.acm.org/10.1145/1148170.1148249.

[98] L. Kelly, Y. Chen, M. Fuller, and G. J. F. Jones, "A study of remembered context for information access from personal digital archives," in *Proceedings of the International Symposium on Information Interaction in Context (IIiX '08)*, New York, NY, USA, pp. 44–50, 2008. ISBN 978-1-60558-310-5. doi: http://doi.acm.org/10.1145/1414694.1414706.

[99] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, "Characterizing Web content, user interests, and search behavior by reading level and topic," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '12)*, New York, NY, USA, pp. 213–222, 2012. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124323. URL http://doi.acm.org/10.1145/2124295.2124323.

[100] S. Kinsella, V. Murdock, and N. O'Hare, ""I'm eating a sandwich in glasgow": modeling locations with tweets," in *Proceedings of the International Workshop on Search and Mining User-generated Contents (SMUC '11)*, New York, NY, USA, pp. 61–68, 2011. ISBN 978-1-4503-0949-3. doi: 10.1145/2065023.2065039. URL http://doi.acm.org/10.1145/2065023.2065039.

[101] J. Kleinberg, "Authorative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, pp. 604–632, 1999.

[102] A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan, "Modeling and analysis of cross-session search tasks," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '11)*, New York, NY, USA, pp. 5–14, 2011. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2009922. URL http://doi.acm.org/10.1145/2009916.2009922.

[103] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar, "Searching with context," in *Proceedings of the International Conference on World Wide Web (WWW '06)*, New York, NY, USA, pp. 477–486, 2006. ISBN 1-59593-323-9. doi: http://doi.acm.org/10.1145/1135777.1135847.

[104] R. Kraft, F. Maghoul, and C. C. Chang, "Y!Q: contextual search at the point of inspiration," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '05)*, New York, NY, USA, pp. 816–823, 2005. ISBN 1-59593-140-6. doi: http://doi.acm.org/10.1145/1099554.1099746.

[105] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais, "Understanding temporal query dynamics," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '11)*, New York, NY, USA, pp. 167–176, 2011. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935862. URL http://doi.acm.org/10.1145/1935826.1935862.

[106] M. Lalmas, "Aggregated search," in *Advanced Topics in Information Retrieval*, (M. Melucci and R. Baeza-Yates, eds.), Springer, 2011.

[107] M. Lalmas and I. Ruthven, "A survey on the use of relevance feedback for information access systems," *Knowledge Engineering Review*, vol. 18, no. 1, 2003.

[108] R. Y. Lau, P. D. Bruza, and D. Song, "Belief revision for adaptive information retrieval," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '04)*, New York, NY, USA, pp. 130–137, 2004. ISBN 1-58113-881-4. doi: http://doi.acm.org/10.1145/1008992.1009017.

[109] R. Y. K. Lau, P. D. Bruza, and D. Song, "Towards a belief-revision-based adaptive and context-sensitive information retrieval system," *ACM Transactions on Information and Systems*, vol. 26, no. 2, pp. 1–38, 2008. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1344411.1344414.

[110] U. Lee, Z. Liu, and J. Cho, "Automatic identification of user goals in Web search," in *Proceedings of the International Conference on World Wide Web (WWW '05)*, New York, NY, USA, pp. 391–400, 2005. ISBN 1-59593-046-9. doi: http://doi.acm.org/10.1145/1060745.1060804. URL http://doi.acm.org/10.1145/1060745.1060804.

[111] R. Lempel and S. Moran, "SALSA: The stochastic approach for link-structure analysis," *ACM Transactions on Information and Systems*, vol. 19, pp. 131–160, 2001.

[112] J. Leveling, "Exploring term selection for geographic blind feedback," in *Proceedings of the ACM Workshop on Geographical Information Retrieval (GIR '07)*, New York, NY, USA, pp. 43–48, 2007. ISBN 978-1-59593-828-2. doi: 10.1145/1316948.1316961. URL http://doi.acm.org/10.1145/1316948.1316961.

[113] L. Levinson, L. Rabiner, and M. Sondhi, "An introduction to the application and theory of probabilistic functions of Markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62-II, pp. 1035–1075, 1983.

[114] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the International Conference on World Wide Web (WWW '10)*, New York, NY, USA, pp. 661–670, 2010. ISBN 978-1-60558-799-8. doi: http://doi.acm.org/10.1145/1772690.1772758.

[115] T. Li, N. Liu, J. Yan, G. Wang, F. Bai, and Z. Chen, "A Markov chain model for integrating behavioral targeting into contextual advertising," in *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD '09)*, New York, NY, USA, pp. 1–9, 2009. ISBN 978-1-60558-671-7. doi: http://doi.acm.org/10.1145/1592748.1592750.

[116] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, and R. Jin, "Exploitation and exploration in a performance based contextual advertising system," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '10)*, New York, NY, USA, pp. 27–36, 2010. ISBN 978-1-4503-0055-1. doi: http://doi.acm.org/10.1145/1835804.1835811.

[117] X. Li, Y.-Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '08)*, New York, NY, USA, pp. 339–346, 2008. ISBN 978-1-60558-164-4. doi: http://doi.acm.org/10.1145/1390334.1390393. URL http://doi.acm.org/10.1145/1390334.1390393.

[118] X. Li, Y.-Y. Wang, D. Shen, and A. Acero, "Learning with click graph for query intent classification," *ACM Transactions on Information and Systems*, vol. 28, pp. 12:1–12:20, July 2010. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1777432.1777435. URL http://doi.acm.org/10.1145/1777432.1777435.

[119] M. D. Lieberman and H. Samet, "Multifaceted toponym recognition for streaming news," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '11)*, New York, NY, USA, pp. 843–852, 2011. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010029. URL http://doi.acm.org/10.1145/2009916.2010029.

[120] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: Topic tracking in tweet streams," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '11)*, New York, NY, USA, pp. 422–429, 2011. ISBN 978-1-4503-0813-7. doi: http://doi.acm.org/10.1145/2020408.2020476. URL http://doi.acm.org/10.1145/2020408.2020476.

[121] F. Liu, C. Yu, and W. Meng, "Personalized Web search by mapping user queries to categories," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '02)*, New York, NY, USA, pp. 558–565, 2002. ISBN 1-58113-492-4. doi: http://doi.acm.org/10.1145/584792.584884.

[122] J. Luxenburger, S. Elbassuoni, and G. Weikum, "Matching task profiles and user needs in personalized Web search," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM '08)*, New York, NY, USA, pp. 689–698, 2008. ISBN 978-1-59593-991-3. doi: http://doi.acm.org/10.1145/1458082.1458175.

[123] Z. Ma, G. Pant, and O. R. L. Sheng, "Interest-based personalized search," *ACM Transactions on Information and Systems*, vol. 25, p. 5, 2007. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1198296.1198301.

[124] M. Melucci, "An evaluation of automatically constructed hypertexts for information retrieval," *Journal of Information Retrieval*, vol. 1, pp. 57–80, 1999.

[125] M. Melucci, "A basis for information retrieval in context," *ACM Transactions on Information and Systems*, vol. 26, pp. 1–41, 2008. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1361684.1361687.

[126] M. Melucci, *Search Engines and Rank Correlation,* volume 4 of *Library and Information Science,* chapter 8. pp. 203–224. Emerald, 2012.

[127] M. Melucci and L. Pretto, "PageRank: When order changes," in *Proceedings of the European Conference on IR Research on Advances in Information Retrieval (ECIR '07)*, Rome, Italy, pp. 581–588, 2007.

[128] M. Melucci and J. Rehder, "Using semantic annotations for automatic hypertext link generation in scientific texts," in *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, (N. Ashish and C. Goble, eds.), Sanibel Island, Florida, USA: Sun SITE Central Europe (CEUR), October 2003. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/.

[129] M. Melucci and R. W. White, "Utilizing a geometry of context for enhanced implicit feedback," in *Proceedings of the ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*, New York, NY, USA, pp. 273–282, 2007. ISBN 978-1-59593-803-9. doi: http://doi.acm.org/10.1145/1321440.1321480.

[130] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '05)*, New York, NY, USA, pp. 472–479, 2005. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076115.

[131] S. Mizzaro, "Relevance: The whole history," *Journal of the American Society for Information Science*, vol. 48, pp. 810–832, 1997.

[132] National Institute of Standards and Technology (NIST), *Text Retrieval Conference WWW site*. http://trec.nist.gov.

[133] K. B. Ng, P. Kantor, T. Strzalkowski, N. Wacholder, R. Tang, B. Bai, R. Rittman, P. Song, and Y. Sun, "Automated judgment of document qualities," *Journal of American Society Information Science and Technology*, vol. 57, pp. 1155–1164, July 2006. ISSN 1532-2882. doi: 10.1002/asi.v57:9. URL http://dx.doi.org/10.1002/asi.v57:9.

[134] Oxford Dictionary, *The New Oxford American Dictionary*. 2005.

[135] E. Palmer, *Graphical Evolution: An Introduction to the Theory of Random Graphs*. Wiley-Interscience, 1985.

[136] E. Peserico and L. Pretto, "HITS can converge slowly, but not too slowly, in score and rank," in *Proceedings of the Annual International Conference on Computing and Combinatorics (COCOON '09)*, Berlin, Heidelberg, pp. 348–357, 2009. ISBN 978-3-642-02881-6. URL http://dx.doi.org/10.1007/978-3-642-02882-3_35.

[137] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search," *Communications of ACM*, vol. 45, pp. 50–55, September 2002. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/567498.567526. URL http://doi.acm.org/10.1145/567498.567526.

[138] J. Ponte and W. Croft, "A language modeling approach to information retrieval," in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, pp. 275–281, August 1998.

[139] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang, "The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the internet," *International Journal of Geographical Information Science*, vol. 21, pp. 717–745, 2007. doi: 10.1080/13658810601169840. URL http://www.tandfonline.com/doi/abs/10.1080/13658810601169840.

[140] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[141] F. Radlinski and T. Joachims, "Query chains: learning to rank from implicit feedback," in *Proceedings of the ACM International Conference on Knowledge Discovery in Data Mining (KDD '05)*, New York, NY, USA, pp. 239–248, 2005. ISBN 1-59593-135-X. doi: 10.1145/1081870.1081899. URL http://doi.acm.org/10.1145/1081870.1081899.

[142] F. Radlinski, R. Kleinberg, and T. Joachims, "Learning diverse rankings with multi-armed bandits," in *Proceedings of the International Conference on Machine Learning (ICML '08)*, New York, NY, USA, pp. 784–791, 2008. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390255. URL http://doi.acm.org/10.1145/1390156.1390255.

[143] J. Raper, "Geographic relevance," *Journal of Documentation*, vol. 63, no. 6, pp. 836–852, 2007.

[144] T. Reichenbacher, "Geographic relevance in mobile services," in *Proceedings of the International Workshop on Location and the Web (LOCWEB '09)*, New York, NY, USA, pp. 10:1–10:4, 2009. ISBN 978-1-60558-457-7. doi: 10.1145/1507136.1507146. URL http://doi.acm.org/10.1145/1507136.1507146.

[145] T. Reichenbacher and S. De Sabbata, "Geographic relevance: different notions of geographies and relevancies," *SIGSPATIAL Special*, vol. 3, pp. 67–70, July 2011. ISSN 1946-7729. doi: 10.1145/2047296.2047310. URL http://doi.acm.org/10.1145/2047296.2047310.

[146] S. Robertson, M. Vojnovic, and I. Weber, "Rethinking the ESP game," in *Proceedings of the International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*, New York, NY, USA, pp. 3937–3942, 2009. ISBN 978-1-60558-247-4. doi: http://doi.acm.org/10.1145/1520340.1520597. URL http://doi.acm.org/10.1145/1520340.1520597.

[147] D. E. Rose and D. Levinson, "Understanding user goals in Web search," in *Proceedings of the International Conference on World Wide Web (WWW '04)*, New York, NY, USA, pp. 13–19, 2004. ISBN 1-58113-844-X. doi: http://doi.acm.org/10.1145/988672.988675. URL http://doi.acm.org/10.1145/988672.988675.

[148] I. Ruthven, "The context of the interface," in *Proceedings of the International Symposium on Information Interaction in Context (IIiX '08)*, New York, NY, USA, pp. 3–5, 2008. ISBN 978-1-60558-310-5. doi: http://doi.acm.org/10.1145/1414694.1414697.

[149] I. Ruthven, "Information retrieval in context," in *Advanced Topics in Information Retrieval,* chapter 8, (M. Melucci and R. Baeza-Yates, eds.), Springer, 2011.

[150] M. Sanderson, "Test collection based evaluation of information retrieval systems," *Foundations and Trends in Information Retrieval*, vol. 4, 2010.

[151] M. Sanderson and Y. Han, "Search words and geography," in *Proceedings of the ACM Workshop on Geographical Information Retrieval (GIR '07)*, New York, NY, USA, pp. 13–14, 2007. ISBN 978-1-59593-828-2. doi: 10.1145/1316948.1316952. URL http://doi.acm.org/10.1145/1316948.1316952.

[152] T. Saracevic, "Relevance: A review of and a framework for thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, pp. 321–343, 1975.

[153] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1915–1933, 2007.

[154] J. G. Shanahan, "Digital advertising: An information scientist's perspective," in *Advanced Topics in Information Retrieval*, (M. Melucci and R. Baeza-Yates, eds.), Berlin, Germany, pp. 209–237, 2011.

[155] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[156] X. Shen, B. Tan, and C. Zhai, "Implicit user modeling for personalized search," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '05)*, New York, NY, USA, pp. 824–831, 2005. ISBN 1-59593-140-6. doi: http://doi.acm.org/10.1145/1099554.1099747.

[157] G. Smith, P. Baudisch, G. Roverson, M. Czerwinski, B. Meyers, D. Robbins, and D. Andrews, "Groupbar: The taskbar evolved," in *Proceedings of OZCHI 2003: New Directions in Interaction, Information Environments, Media and Technology*, pp. 34–43, 2003.

[158] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck, "Probabilistic models for personalizing Web search," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '12)*, New York, NY, USA, pp. 433–442, 2012. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124348. URL http://doi.acm.org/10.1145/2124295.2124348.

[159] A. Spink, *Information Behavior: An Evolutionary Instinct*. Springer, 1st ed., 2010. ISBN 3642114962, 9783642114960.

[160] J. Teevan, S. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR '05)*, New York, NY, USA: ACM Press, pp. 449–456, 2005. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076111.

[161] J. Teevan, S. T. Dumais, and E. Horvitz, "Potential for personalization," *ACM Transactions on Computer-Human Interactions*, vol. 17, pp. 1–31, 2010. ISSN 1073-0516. doi: http://doi.acm.org/10.1145/1721831.1721835.

[162] J. Teevan, S. T. Dumais, and D. J. Liebling, "To personalize or not to personalize: Modeling queries with variation in user intent," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '08)*, New York, NY, USA, pp. 163–170, 2008. ISBN 978-1-60558-164-4. doi: http://doi.acm.org/10.1145/1390334.1390364.

[163] J. Teevan, D. J. Liebling, and G. Ravichandran Geetha, "Understanding and predicting personal navigation," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '11)*, New York, NY, USA, pp. 85–94, 2011. ISBN 978-1-4503-0493-1. doi: http://doi.acm.org/10.1145/1935826.1935848. URL http://doi.acm.org/10.1145/1935826.1935848.

[164] J. Teevan, M. R. Morris, and S. Bush, "Discovering and using groups to improve personalized search," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '09)*, New York, NY, USA, pp. 15–24, 2009. ISBN 978-1-60558-390-7. doi: 10.1145/1498759.1498786. URL http://doi.acm.org/10.1145/1498759.1498786.

[165] K. Tuite, D.-Y. H. N. Snavely, N. Tabing, and Z. Popovic, "Photocity: Training experts at large-scale image acquisition through a competitive game," in *Proceedings of the Annual Conference on Human Factors in Computing Systems (CHI '11)*, New York, NY, USA, pp. 1383–1392, 2011. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979146. URL http://doi.acm.org/10.1145/1978942.1979146.

[166] D. Tunkelang, *Faceted Search*. Morgan and Claypool, 2009.

[167] C. van Rijsbergen, *Information Retrieval,* chapter 6, pp. 144–183. London: Butterworths, second ed., 1979.

[168] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '04)*, New York, NY, USA, pp. 319–326, 2004. ISBN 1-58113-702-8. doi: http://doi.acm.org/10.1145/985692.985733. URL http://doi.acm.org/10.1145/985692.985733.

[169] R. W. White, "Using searcher simulations to redesign a polyrepresentative implicit feedback interface," *Information Processing Management*, vol. 42, pp. 1185–1202, 2006. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2006.02.005.

[170] R. W. White, P. Bailey, and L. Chen, "Predicting user interests from contextual information," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '09)*, New York, NY, USA, pp. 363–370, 2009. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1572005.

[171] R. W. White, M. Bilenko, and S. Cucerzan, "Studying the use of popular destinations to enhance Web search interaction," in *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '07)*, New York, NY, USA, pp. 159–166, 2007. ISBN 978-1-59593-597-7. doi: http://doi.acm.org/10.1145/1277741.1277771. URL http://doi.acm.org/10.1145/1277741.1277771.

[172] R. W. White, J. M. Jose, and I. Ruthven, "A task-oriented study on the influencing effects of query-biased summarisation in Web searching," *Information Processings Management*, vol. 39, pp. 707–733, 2003. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(02)00033-X.

[173] R. W. White and D. Kelly, "A study on the effects of personalization and task information on implicit feedback performance," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '06)*, New York, NY, USA, pp. 297–306, 2006. ISBN 1-59593-433-2. doi: http://doi.acm.org/10.1145/1183614.1183659.

[174] R. W. White and R. A. Roth, *Exploratory Search: Beyond the Query-response Paradigm*. Morgan and Claypool, 2009.

[175] Wikipedia, *Definition of Dendrogram*. http://en.wikipedia.org/wiki/Dendrogram, October 2011.

[176] X. Yi, H. Raghavan, and C. Leggetter, "Discovering users' specific geo-intention in Web search," in *Proceedings of the International Conference on World Wide Web (WWW '09)*, New York, NY, USA, pp. 481–490, 2009. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526774. URL http://doi.acm.org/10.1145/1526709.1526774.

[177] E. Yom-Tov and F. Diaz, "Out of sight, not out of mind: On the effect of social and physical detachment on information need," in *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR '11)*, New York, NY, USA, pp. 385–394, 2011. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2009970. URL http://doi.acm.org/10.1145/2009916.2009970.

[178] Y. Yue and T. Joachims, "Predicting diverse subsets using structural svms," in *Proceedings of the International Conference on Machine Learning (ICML '08)*, New York, NY, USA, pp. 1224–1231, 2008. ISBN 978-1-60558-205-4. doi: http://doi.acm.org/10.1145/1390156.1390310.

[179] Y. Yue and T. Joachims, "Interactively optimizing information retrieval systems as a dueling bandits problem," in *Proceedings of the Annual International Conference on Machine Learning (ICML '09)*, New York, NY, USA, pp. 1201–1208, 2009. ISBN 978-1-60558-516-1. doi: http://doi.acm.org/10.1145/1553374.1553527.

[180] Y. Zhou and W. B. Croft, "Document quality models for Web ad-hoc retrieval," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '05)*, New York, NY, USA, pp. 331–332, 2005. ISBN 1-59593-140-6. doi: 10.1145/1099554.1099652. URL http://doi.acm.org/10.1145/1099554.1099652.

[181] G. Zhu and G. Mishne, "Mining rich session context to improve Web search," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD '09)*, New York, NY, USA, pp. 1037–1046, 2009. ISBN 978-1-60558-495-9. doi: http://doi.acm.org/10.1145/1557019.1557131.

[182] G. Zhu and G. Mishne, "Clickrank: Learning session-sontext models to enrich Web search ranking," *ACM Transactions on Web*, vol. 6, pp. 1:1–1:22, March 2012. ISSN 1559-1131. doi: 10.1145/2109205.2109206. URL http://doi.acm.org/10.1145/2109205.2109206.

[183] Z. Zhuang, C. Brunk, and C. L. Giles, "Modeling and visualizing geo-sensitive queries based on user clicks," in *Proceedings of the International Workshop on Location and the Web (LOCWEB '08)*, New York, NY, USA, pp. 73–76, 2008. ISBN 978-1-60558-160-6. doi: 10.1145/1367798.1367811. URL http://doi.acm.org/10.1145/1367798.1367811.