Semantic search log analysis: a method and a study on professional image search[*]

Vera Hollink, Theodora Tsikrika, Arjen P. de Vries

Centrum Wiskunde en Informatica
Science Park 123, 1098 XG Amsterdam, The Netherlands

V.Hollink@cwi.nl, theodora.tsikrika@acm.org, arjen@acm.org

Abstract

Existing methods for automatically analyzing search logs describe search behavior on the basis of purely syntactic differences (overlapping terms) between queries. Although these statistics at a syntactic level provide valuable insights into the complexity and successfulness of search interactions, they offer a limited interpretation of the observed searching behavior, as they do not consider the *semantics* of users' queries. Recently, large amounts of semantic information have become publicly available in the form of *linked data*. In this paper we propose a method to exploit this information to enrich search queries with linked data entities so as to determine the semantic types of the queries and the relations between queries that are consecutively entered in a search session.

This work provides also an in-depth analysis of the search logs of the commercial picture portal of a European news agency, which offers access to photographic images to professional users. Compared to previous image search log analyses, in particular those of professional users, we consider a much larger dataset. We analyze the logs both in the more traditional syntactic way and using the newly proposed semantic approach, and compare the results. Our findings show the benefits of using semantics for search log analysis: the identified types of query modifications cannot be appropriately analyzed with a purely statistical approach that only considers term overlap, since queries related in the most frequent ways do not usually share terms. We discuss implications of our findings for improving log analysis, image collection management, and search engine design.

1 Introduction

Users interact with web search engines in order to satisfy their information needs. The examination of the users' information searching behavior aims at providing an understanding of the usage of such systems; the obtained information can then be ultimately used for improving user experience and system effectiveness. In particular, the findings of such investigations can aid the development of user interfaces and mechanisms that assist users during search, and also benefit system design, content management, and research into the development of more effective retrieval approaches.

The study of information searching behavior is typically conducted with research methods drawn from Social Sciences, including methods such as surveys and laboratory-based user-centered experiments (Ingwersen and Järvelin, 2005). For studying the behavior of web search engines' users, one of the most effective methods is the analysis of their search interaction data collected in these engines' transaction logs (Jansen et al., 2008a). Such logging mechanisms are able to unobtrusively record large amounts of the entire range of user-system interactions of a sizable number of users, in naturalistic settings, over significant time periods (Jansen et al., 2008b). On the other hand, though, search logs do not record the underlying context and situation of users' searching process, or any other qualitative user aspects, and therefore, the users' information needs, decision processes, and satisfaction with the system remain implicit (Jansen, 2006, 2008a). Despite these inherent limitations, search logs have been widely examined in order to gain an understanding of web users' searching behavior (e.g., Lau and Horvitz, 1999, Jansen et al., 2000b,a, Goodrum and Spink, 2001, Spink et al., 2001, Jansen et al., 2004, Jansen and Spink, 2006, Rieh and Xie, 2006, Whittle et al., 2007, Bozzon et al., 2007, Costa and Seco, 2008, Jansen et al., 2009, Tjondronegoro et al., 2009, Boldi et al., 2009, Huang and Efthimiadis, 2009).

Search log analysis examines the characteristics of users' searching interactions in order both to describe the observed trends and patterns, and to make predictive inferences (Jansen, 2006, 2008a). Typically, the interactions that are examined include query submission, query modification (i.e., the successive submission of queries by a user engaging in an iterative interactive process in the context of a single session), and results viewing, with the analysis taking place at three levels of granularity (Jansen, 2006): *terms*, *queries*, and *sessions* (i.e., series of searching interactions performed within a given temporal span by a single searcher).

Current approaches to search log analysis (Jansen et al., 2008a) describe the main features of the observed interactions in purely statistical terms by reporting, for instance, in some aggregate form, the number of terms per query, the number of queries per session, the number of results viewed, and the differences in the terms occurring in consecutive queries submitted by the same user. Statistics collected at the level of the query or the terms it is composed of provide valuable insights into the complexity and successfulness of search interactions, and into the users' behavioral patterns regarding query formulation

2

and modification. However, this syntactic analysis at the level of term overlap offers only a partial interpretation of the observed search behavior since it does not consider the *semantics* of the available information. Intuitively, taking into account the semantics of individual queries, as well as the semantic relations between queries, would allow a finer-grained and potentially more meaningful analysis of search log data; this, combined with a statistical analysis, may offer a new understanding of user behavior not visible from usage statistics alone.

To this end, this work proposes a novel *semantic search log analysis* method that enriches current approaches by exploiting the knowledge in a *linked data cloud* (Berners-Lee, 2006, Bizer et al., 2009), i.e., a collection of interlinked RDF triples (Klyne and Carroll, 2004) that encode what is predicated of specific entities, e.g., that David Beckham is a football player and married to Victoria Beckham. The proposed method adds a semantic dimension to search log analysis by mapping the submitted queries onto *entities* in the linked data, thus enabling the identification of query types and semantic relations between queries. Consider, for instance, the queries `David Beckham` and `Joe Cole` submitted in succession by a single user. An analysis based on term overlap cannot identify a relation between these two queries. A semantic analysis would treat these as entities, rather than as mere strings, and would determine that each of these queries corresponds to a footballer (the semantic type of the queries) and that their relation is that both play for the English national football team (the semantic type of the query modification). Based on the premise that users' searching behavior, as this is recorded in search logs, reflects aspects of the situation and context in which this behavior occurs, the above example indicates that semantic search log analysis has the potential to uncover more evidence regarding these user aspects compared to purely statistical approaches.

Aside the new method, this work also reports a study of the search logs of the commercial picture portal of a European news agency. The picture portal offers access to millions of photographic images to professional users, such as journalists, PR agencies, and editorial staff in newspapers and magazines. Compared to the number of investigations performed on logs of text search engines, only few studies have looked into image search logs (Goodrum and Spink, 2001, Jansen et al., 2000a, Özmutlu et al., 2003, Jansen, 2008b, Jansen et al., 2004, Tjondronegoro et al., 2009). And, in a professional context, search log analysis has only been applied to small scale samples (Markkula and Sormunen, 2000, Jörgensen and Jörgensen, 2005, Westman and Oittinen, 2006). The study reported here analyzes the image search logs collected by the picture portal over a 10-month period, comparing the results of a statistical term-based analysis with those obtained by applying the proposed semantic approach. Particular emphasis is given on the application of the proposed method for analyzing the query modification behavior of users and thus uncovering the most prominent modification patterns.

1.1 Research objectives

The two main questions addressed in this paper are:

1. How can we enrich search log analysis by intoducing a semantic dimension, i.e., how can we semantically analyze a search log to discover what types of queries are posed, how queries are modified, and what the differences are between successful and unsuccessful queries?
   Existing methods for analyzing logs do not use the semantics of the queries. Including semantic knowledge in the analysis can potentially provide further insights in the searching behavior.

2. What are the search characteristics of professional users searching for images in a journalistic context?
   Image search behavior of professional users has not yet been investigated in sufficient depth. In the few studies that addressed this topic very small log samples were analyzed.

To answer the first research question we developed a novel search log analysis method that exploits semantic knowledge. To analyze the semantics of queries and query modifications, it does not suffice to treat queries as mere strings. Instead, we need to augment queries with semantic information by matching them to entities with well-defined properties and semantic relations to other entities. Fortunately, large sources of semantic information have recently become available in the form of linked data. Combining search log analysis with linked data poses several challenges. 1) How can we match queries to entities in linked data? 2) How can we exploit the properties and relations of entities to reveal meaningful information about queries and query pairs? 3) How can we use the statistics of the properties and relations of the entities matching the queries in the search logs to identify patterns in search behavior that hold for a large number of users? In this paper we present solutions for each of these challenges. We explain how our method can be used to analyze the types of the users' queries at multiple levels and to discover the users' query modification strategies. To assess the added value of the semantic search log analysis, we compare the findings of our approach to the results of existing syntactic methods.

The second question is answered by analyzing a large scale image search log of professional users with both syntactic methods and the presented semantic method. Firstly, we explore the types of queries these users posed and examine differences between the types of queries that lead to a click on an image (successful queries) and the types of queries that did not lead to a click on an image (unsuccessful queries). Secondly, we examine the strategies the users followed to modify queries. We analyze commonly used modification patterns and explore the differences between modifications after successful queries and unsuccessful queries. We discover that two classes of patterns occur frequently and compute the overlap of these classes with modification types found through syntactic analysis. Finally, we discuss the implications of our findings for various applications, including search support, as well as management and annotation of journalistic image collections.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 presents the semantic search log analysis method. Sec-

tion 4 describes the search log data that are studied and presents some of their statistics. The proposed semantic approach is then applied on these data and compared to syntactic approaches, with Section 5 reporting on query types, and Section 6 on query modification patterns. Section 7 discusses implications of our findings and Section 8 concludes by summarizing the main findings and outlining directions for future work.

2   Related work

Search log analysis is typically performed at three levels of granularity (Jansen, 2006): the session level, the query level, and the term level. The session level of analysis provides insights in the duration, the complexity and the successfulness of search sessions. Query level analysis focuses on the number of terms in the users' queries and on query modifications, examining how often and in what way users modify their queries. Term level analysis examines the frequency distribution of terms in queries and the topics of the terms. Previous search log analyses conducted at these three levels are reviewed in Section 2.1, whereas Section 2.2 zooms in on query modification. Given that this work reports on the analysis of the search logs of a commercial picture portal, Section 2.1 focuses only on studies of image search logs. Section 2.2, though, also discusses studies on non-image search, since very few studies have examined query modifications in image search.

2.1   Image search log analysis

Most studies of image search logs have examined log files of general-purpose web search engines with a varied user population. As shown in Table 1, Altavista has been studied in Jansen et al. (2004), Excite in Goodrum and Spink (2001), Jansen et al. (2000a), Özmutlu et al. (2003), and Jansen (2008b), and the meta-search engine Dogpile in Tjondronegoro et al. (2009). Exceptions are the studies by Jörgensen and Jörgensen (2005), who analyzed the logs of professional users subscribed to the services of a commercial image provider, and by Markkula and Sormunen (2000) and Westman and Oittinen (2006), who examined the logs of journalists and archivists searching in a digital newspaper photo archive. In these studies of professional image search (Markkula and Sormunen, 2000, Jörgensen and Jörgensen, 2005, Westman and Oittinen, 2006) the collection in which the users searched consisted only of images. The Altavista and Dogpile search engines examined in Jansen et al. (2004) and Tjondronegoro et al. (2009) provided access to image, audio, video, and textual content, where users could distinguish between these four collections by selecting the corresponding tab. Jansen et al. (2004) compare searches in all four collections, whereas Tjondronegoro et al. (2009) focus on searches in the three non-textual collections. At the time of study of the Excite search logs, the search engine offered only one search box which mixed text, image, audio, and video results. Goodrum and Spink (2001) and Jansen (2008b) identified queries for image material in

Table 1: Previous studies on image search log data: the search engine providing the logs for analysis, the method of analysis, the year in which the logs were collected, and the number of image queries and sessions that were analyzed.

Study	Search engine	Analysis	Year	#Queries	#Sessions
general-purpose web image search					
Jansen et al. (2000a)	Excite	automatic	1997	27,144	8,310
Goodrum & Spink (2001)	Excite	automatic	1997	33,149	9,855
Özmutlu et al. (2003)	Excite	automatic	2001	53,985	16,670
Jansen (2008b)	Excite	manual	2001	587	unknown
Jansen et al. (2004)	AltaVista	automatic	2002	127,614	26,720
Tjondronegoro at al. (2009)	Dogpile	automatic	2006	562,380	196,755
professional image search					
Markkula & Sormunen (2000)	Newspaper archive	manual	1996-97	unknown	27
Westman & Oittinen (2006)	Newspaper archive	manual	2004	1788	unknown
Jörgensen & Jörgensen (2005)	Commercial image provider	manual	2005	2810	841

the Excite search logs by extracting queries that contained one of 28 predefined image related keywords (e.g., `image, pics, photo`). In the same way, Jansen et al. (2000a) and Özmutlu et al. (2003) identified in these logs image, audio, and video queries.

In half of these studies, the log files were analyzed automatically (Jansen et al., 2004, Tjondronegoro et al., 2009, Goodrum and Spink, 2001, Jansen et al., 2000a, Özmutlu et al., 2003). Automatic processing allows for the analysis of a large sample of the search logs, ranging from 8,310 image search sessions in Jansen et al. (2000a) to 196,755 sessions in Tjondronegoro et al. (2009) (see Table 1). In contrast, Jörgensen and Jörgensen (2005) manually coded a much smaller log sample of 841 sessions, while Markkula and Sormunen's study (2000) was limited to 27 sessions only.

Most studies that compared search in various types of collections found that users search much more often for images than for audio and video content, but less frequently than for general content (Jansen et al., 2004, Tjondronegoro et al., 2009, Jansen et al., 2000a). Image queries occurred between 3.6 and 4.6 times more often than video queries and between 2.2 and 8.4 times as much as audio queries. Only in Özmutlu et al. (2003), image queries occurred less frequently than audio queries, but only in one of the three examined years of log data. In total, image queries made up 1.7% (Özmutlu et al., 2003) to 5.5% (Jansen et al., 2004) of the queries entered in general-purpose web search engines.

Jansen et al. (2004) and Özmutlu et al. (2003) found that image sessions comprise on average more queries (4.8) than searches for other types of content and that image queries contain more terms (4.0). From this they conclude that image search is a more complex cognitive task than other types of search. The larger number of queries per session is confirmed by Jansen et al. (2000a) and Tjondronegoro et al. (2009). However, these studies both find that the average query length of image queries is not longer than in other collections. The average number of queries per session appears to be fairly constant across the various general-purpose web image search engines: between 2.8 and 4.8. The number

of terms per query is also similar for all search engines: 2.3 to 4.0. Smaller numbers of terms are found for journalists searching for images professionally: 1.48 in Westman and Oittinen (2006) and 1.9 in Jörgensen and Jörgensen (2005) (although in this study bound phrases such as 'New York' were counted as one term).

Few studies have analyzed the browse and click behavior of users after they had entered a query. In the study of Tjondronegoro et al. (2009), 56% of the image *sessions* resulted in a click on one of the search results, a much lower percentage than for audio and video queries. A comparable result for professional users has been found in Jörgensen and Jörgensen (2005), where 26% of the *queries* resulted in clicking or downloading an image. Jansen et al. (2004) examined the number of result pages that were opened by users of Altavista and found that these users looked at very few result pages: in only 37% of the cases more than one page was viewed. A very different result was found in Jörgensen and Jörgensen (2005), who report that after 90.1% of the queries of professional users more than one page was viewed. Markkula and Sormunen (2000) also found that professionals use much browsing. Based on this result, combined with the shorter queries of these users, they hypothesize that journalists often use broad and short queries to make sure no relevant images are missed and rely on browsing to find the best images among the returned results.

So far we discussed research on how users search, but not on what they search for. Various methods have been developed to automatically classify the topics and types of search queries. For example, Beitzel et al. (2007) classify queries into predefined topics by comparing the queries to manually classified queries and learning selectional preferences from unlabeled queries. Shen et al. (2006) apply text classification methods on snippets and categories of search results of the queries. An extensive overview of query classification methods can be found in Özmutlu et al. (2009). Although such query classification methods have been used to analyze the topics of text search queries (e.g., Bar-Ilan et al. (2009) for MSN queries), few studies have applied them to image search logs. An exception is the work by Tjondronegoro et al. (2009), who examined the topics of 2818 queries, using OpenCalais (Ope, 2010), a public Web Service for the classification of natural language text into the semantic categories defined by Calais (a Thomson Reuters company). In a number of other papers (Jansen et al., 2004, Goodrum and Spink, 2001, Jansen, 2008b, Westman and Oittinen, 2006, Markkula and Sormunen, 2000, Jörgensen and Jörgensen, 2005), samples of (frequently occurring) image queries have been classified by hand. Consistently, these studies showed that image queries cover a wide variety of topics, judging from the large number of queries that occur only once (Jansen et al., 2004, Goodrum and Spink, 2001). In the general-purpose web search engines many queries refer to adult content (Jansen et al., 2004, Goodrum and Spink, 2001, Tjondronegoro et al., 2009, Jansen, 2008b). Other common topics were medical conditions and people (Tjondronegoro et al., 2009). Jansen (2008b) found that most queries (82%) asked for general concepts (rather than named entities). Furthermore, people and people-related queries occurred often. In logs of professional users, the main topics were business, computers, people,

activities and sports, whereas queries for adult content are seen much less frequently (Jörgensen and Jörgensen, 2005). Queries of journalists consist mainly of named entities (56% in both Westman and Oittinen (2006) and (Markkula and Sormunen, 2000)), in particular persons' names (40% in Westman and Oittinen (2006)). General concepts occur less frequently: 44% in Westman and Oittinen (2006) and 18% in Markkula and Sormunen (2000).

In conclusion, the behavior of users who search for images in general-purpose web search engines is reported upon in various papers. These papers have shown that image search is a complex task, which involves many queries and often does not lead to results that the users find worth clicking. Much less research has addressed the behavior of professional users. Exceptions are the studies of Markulla and Sormunen (2000), Jörgensen and Jörgensen (2005), and Westman and Oittinen (2006). These studies provide a thorough analysis of professional search, but are limited to very small numbers of search sessions. In our paper we examine professional image search by analyzing the search interactions of a much larger number of professional users by 'automating' the coding, exploiting the opportunities provided by the Linked Open Data initiative.

2.2 Query modification

Research on query modification studies pairs of queries that are successively submitted in a search session. Successive query pairs are classified into a number of modification types and the use of these types is analyzed. Below we summarize previous research on query modification. We discuss the source of the search logs that are examined, the modification types that are defined and the main conclusions about these types. Statistics of the discussed studies are given in Table 2.

The large majority of the studies on query modification used search logs of search engines for textual content. Probably the most studied search logs are those of the Excite search engine (Rieh and Xie, 2006, Whittle et al., 2007, Lau and Horvitz, 1999, Bozzon et al., 2007, Özmutlu, 2009). Other studies analyzed logs of Dogpile (Jansen et al., 2009), Tumba (Costa and Seco, 2008), AOL (Huang and Efthimiadis, 2009), Fast (Özmutlu, 2009), and Yahoo! UK and Yahoo! US (Boldi et al., 2009). Jones and Fain (2003) used client-side logging software to collect logs of a group of users accessing a variety of search engines. He et al. (2002) examined the search logs from an intranet. Bruza and Dennis (1997) used logs from the BabyOIL prototype which gave access to seven databases containing various kinds of documents. In Efthimiadis (2000) query modifications are not studied from search logs but in a laboratory setting involving users performing search tasks. To our knowledge, the only study that examined modifications in image search is the study by Jörgensen and Jörgensen (2005).

In most of these studies, the modification types that are recognized are *specification*, *generalization* and *reformulation*. Observed query modifications are classified either manually (Bruza and Dennis, 1997, Lau and Horvitz, 1999, Jörgensen and Jörgensen, 2005, Rieh and Xie, 2006, Jones and Fain, 2003,

Table 2: Previous studies on search query modification: the search engine providing the logs for analysis, the method of the query modification analysis, the year in which the logs were collected, and the number of queries and sessions that were analyzed.

Study	Search engine	Analysis	Year	#Queries	#Sessions
Lau and Horvitz (1999)	Excite	manual	1997	4,690	unknown
Rieh and Xie (2006)	Excite	manual	2000	2,109	313
Whittle et al. (2007)	Excite	automatic	2001	1,025,838	266,598
Bozzon et al. (2007)	Excite	automatic	unknown	2,007,897	319,566
Özmutlu (2009)	Excite	automatic	1999	5,010	526
Özmutlu (2009)	Fast	automatic	2001	4,997	437
Costa and Seco (2008)	Tumba	automatic	2003	1,700,000	75,320
Jansen et al. (2009)	Dogpile	automatic	2005	1,523,793	964,780
Huang and Efthimiadis (2009)	AOL	automatic	2006	36,389,567	16,069,421
Boldi et al. (2009)	Yahoo!	automatic	2008	17,000,000	7,000,000
Bruza and Dennis (1997)	BabyOIL	manual	1996	4,064	unknown
He et al. (2002)	Intranet	automatic	1999	9,534	1,357
Jones and Fain (2003)	Various	manual	2002	1,236	unknown
Efthimiadis (2000)	INSPEC	manual	unknown	unknown	25
Jörgensen and Jörgensen (2005)	Commercial image provider	manual	2005	2,810	841

Efthimiadis, 2000) or automatically (He et al., 2002, Whittle et al., 2007, Jansen et al., 2009, Boldi et al., 2009, Costa and Seco, 2008, Bozzon et al., 2007, Huang and Efthimiadis, 2009, Özmutlu, 2009). Studies that employ automatic methods usually classify query modifications solely on the basis of terms in the queries. These studies (as well as some of the manual studies) examine whether terms have been added, eliminated or substituted compared to the user's previous query (Jansen et al., 2009, Whittle et al., 2007, Jörgensen and Jörgensen, 2005, He et al., 2002, Bruza and Dennis, 1997, Jones and Fain, 2003, Costa and Seco, 2008). When terms are added, the modification is classified as specification (e.g., from query `Beckham` to query `Beckham Milan`), when terms are eliminated it is classified as generalization (e.g., from `Beckham Milan` to `Beckham`), and when terms are substituted it is classified as reformulation (e.g., from `Beckham Milan` to `Beckham Madrid`). Some of the manual studies do not only look at terms but also classify modifications based on the meaning of the queries (Rieh and Xie, 2006, Jörgensen and Jörgensen, 2005, Lau and Horvitz, 1999, Jones and Fain, 2003, Efthimiadis, 2000). The same main classes are used, but the semantics of the queries are also taken into account during classification. For example, a modification from `Dog` to `Labrador` is classified as a specification, even though the queries do not have any terms in common. Huang and Efthimiadis (2009) use strictly term-based definitions for specifications and generalizations, but for reformulations they require that the substituted terms are semantically related according to WordNet (Fellbaum, 1998), in particular that the term being added is a synonym, hyponym, hypernym, meronym, or holonym of the term being removed. Another intermediate approach is presented by Boldi et al. (2009) who aim at a semantic classification of query modifications into specification, generalization and reformulation by looking at the overlap in query terms, time intervals between queries and features of the user session as a whole. Very different definitions of specification, generalization and reformulation are given by

Bozzon et al. (2007), who use the frequency of the query terms in the document collection. If the terms in the first query are more frequent than the terms in the second query, the modification is called a specification. The opposite case is a generalization. Reformulation is defined as two queries with almost equal term frequencies.

Not all query modifications can be assigned to one of the three modification classes and the proportion of relations that can be recognized varies across the different studies. Using the terms and semantics of the queries, Jones and Fain (2003) are able to manually classify 47% of the modifications, while 66% of the modifications can be classified in Efthimiadis (2000). Based on term overlap alone, Jones and Fain (2003) find a class for 23% of the modifications, He et al. (2002) for 43% of the modifications, and Özmutlu (2009) for 33% of the Fast modifications and 33% of the Excite modifications. Most likely, these differences arise from differences between the search engines and differences in preprocessing steps, such as stemming and session identification. In fact, in the majority of related work unclassified query modifications are assumed to start new search sessions and are not separately reported.

The large majority of the studies find that the most frequently used modification type is reformulation, followed by specification and generalization (Bruza and Dennis, 1997, Lau and Horvitz, 1999, Jörgensen and Jörgensen, 2005, Rieh and Xie, 2006, Whittle et al., 2007, Jansen et al., 2009, Boldi et al., 2009, Costa and Seco, 2008, Efthimiadis, 2000, Özmutlu, 2009). Reformulations occur roughly twice as often as specifications which occur twice as often as generalizations. A noteworthy observation is that there is no difference in this respect between manual and automatic methods or between purely term-based and semantic methods. The only comparable studies in which different proportions are found are He et al. (2002) and Jones and Fain (2003): they find almost equal numbers of reformulations and specifications and a much smaller number of generalizations. No explanation is given for this deviation. An entirely different distribution is found in Bozzon et al. (2007), but this work does not use term overlap between queries, but term frequencies in the document collection.

Sequences of query modifications are examined in Whittle et al. (2007), Jansen et al. (2009), and Boldi et al. (2009). Whittle et al. (2007) found that users tend to repeat the same modification type (e.g., a specification is often followed by another specification). This is not confirmed by Boldi et al. (2009), who found that specifications are usually followed by generalizations and generalizations by specifications. These are also the dominant patterns in Jansen et al. (2009).

Two studies examined the relation between modification types and the time interval between submitting two consecutive queries. Lau and Horvitz (1999) found that specification is most likely after an interval of 20 to 30 seconds while reformulation peaks when the interval is longer than 5 minutes. However, the differences are small. In Huang and Efthimiadis (2009) reformulations also had the longest average time interval (73 seconds), closely followed by generalizations (68 seconds) and specifications (63 seconds).

The relation between modification types and clicks on search results is stud-

ied in Huang and Efthimiadis (2009). They found that generalization and reformulation often occur when the previous query has led to at least one click on a search result, which indicates that these modifications are mainly used after successful queries. Some types of term variations mainly occur when the previous query has not led to a click, suggesting a second attempt to find the same information. Specifications and reformulations appeared to be most successful: these modifications most often resulted in a click.

Besides the three main modification types, sometimes other types are examined as well. A number of studies have looked at the number of times users enter term variations (Rieh and Xie, 2006, Bruza and Dennis, 1997, Jones and Fain, 2003). Such variations include, for instance, modifications from singular to plural forms or vice versa. Term variations are less common than the main modification classes, but they still make up a significant proportion of the query modifications, occurring about half as frequently as generalizations. Bozzon et al. (2007) examined differences between lexical categories (part-of-speech tags) of queries and modified queries. They found that most queries consist of only nouns or combinations of nouns and verbs. Noun queries are most often modified into other noun queries and noun-verb queries into other noun-verb queries. However, transitions between these two lexical types also occur frequently.

In summary, the three term-based modifications (specification, generalization, reformulation) have been extensively studied. In some papers also other term-based modifications are examined, such as term variations and lexical categories. Different authors have researched different variants and aspects of term-based modifications. However, to our knowledge there are no papers that classify query modifications into finer grained, semantic categories.

3   Semantic search log analysis

In this section we present our method for semantic analysis of search logs. We start with a discussion on the main principles of linked data and introduce the linked data sources that are used in our analysis. Then, we explain how we match queries to linked data entities and how the matching entities are used to determine the semantic types of queries and to identify semantic query modification patterns. We end this section with a discussion on the accuracy of the presented method. Figure 1 summarizes the various steps of the method.

3.1   Linked data

Here we briefly review the main concepts of linked data; for an extensive overview see Bizer et al. (2009). The idea of linked data was first described by Tim Berners-Lee (2006) in the form of four principles that prescribe how data should be published on the web. (1) Each entity in the data is referred to by a unique URI. (2) Information about the entity can be attained by looking up the URI via HTTP. Information about (3) entities and (4) links between entities is coded in RDF (Klyne and Carroll, 2004): a set of triples <subject, predicate,

Input: – list of search sessions
             – RDF triples in the used linked data sources
Output: – list of query types and their relative frequencies
             – list of query modification patterns and their support and confidence values

---

Step 1 (Section 3.2, Appendix A) :
             – Map queries in search sessions to entities in RDF triples
Step 2 (Section 3.3, Appendix B) :
             – Determine the types of the entities to which the queries are mapped
             – Count occurrence frequencies of identified types
Step 3 (Section 3.4, Appendix C) :
             – Determine the semantic relations (modifications) between the entities of
               pairs of consecutive queries
             – Count occurrence frequencies of identified modifications
             – Compute support and confidence values of identified modifications

Figure 1: Overview of the method for semantic search log analysis.

object>, where the subject and the predicate are both URIs and the object can either be a URI or a string literal. Following these principles ensures that the data can be easily shared with others, read by both humans and machines, and linked to data from other sources. Examples of RDF triples are given in Figure 2. The first triple provides information about the entity `DBpedia:David Beckham`: it provides a human readable label for the entity. The second triple tells us that David Beckham is of type football player. The third triple states that two URIs from different sources refer to the same concept.

The linked data that we use consists of various interlinked sources: the DBpedia Ontology (DBp, 2009), WordNet 2.0 (Fellbaum, 1998, Van Assem et al., 2006), the Cornetto Lexical Knowledge Base version 1.2 (Vossen et al., 2008), the Getty Thesaurus of Geographical Names (TGN) (The Getty Foundation, 2004b), and the Getty Art and Architecture Thesaurus (AAT) (The Getty Foundation, 2004a). Together these collections comprise 22 million RDF triples. Most of these sources are part of the linked data that is published in the Linking Open Data Project (Lin, 2010) and publicly available on the web.

```
Subject:    http://dbpedia.org/resource/David_Beckham
Predicate:  http://www.w3.org/2000/01/rdf-schema#label
Object:     'David Beckham'

Subject:    http://dbpedia.org/resource/David_Beckham
Predicate:  http://www.w3.org/1999/02/22-rdf-syntax-ns#type
Object:     http://dbpedia.org/ontology/FootballPlayer

Subject:    http://e-culture.multimedian.nl/ns/cornetto/wordnet/d_n-10557
Predicate:  http://e-culture.multimedian.nl/ns/cornetto/wordnet/eqNearSynonym
Object:     http://www.w3.org/2006/03/wn/wn20/instances/synset-bullterrier-noun-1
```
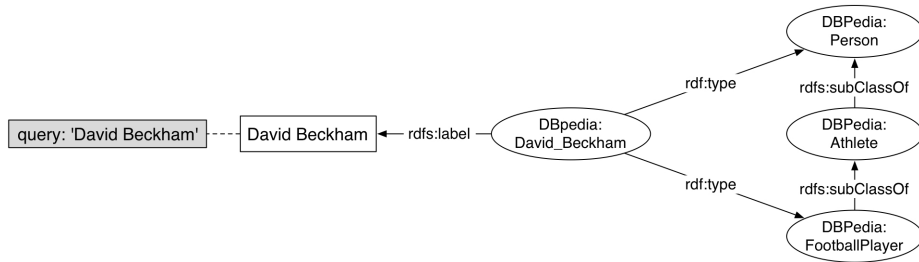
Figure 2: Examples of RDF triples.

Figure 3: Example application of our method for finding query types: the type of query `David Beckham` is at the lowest level `DBpedia:FootballPlayer` and at the highest level `DBpedia:Person`.

3.2 Mapping queries to linked data entities

The first step of our method is to map the queries in the user sessions on entities in the linked data. To map the queries we use the `rdfs:label` property of the entities, which provides a human readable description for the entities (Bri, 2004). An example of an entity and its label are given in Figure 2 (first triple).

We map queries on entities that have an `rdfs:label` that exactly matches the query. If no exact match can be found, the queries are stemmed with the Porter stemmer (Porter, 1980) and mapped onto entities with labels that contain all stemmed query terms. With this method each query is mapped onto zero, one, or multiple entities. We purposely chose a conservative mapping method, sacrificing recall for precision, to reduce the amount of noise in our results. The complete algorithm for mapping queries to entities can be found in Appendix A.

3.3 Identifying query types

To identify the types of queries that users have entered in a search engine, we determine the classes of the entities to which the queries are mapped, as shown in Figure 3. We make use of the `rdf:type` property in the linked data, which is the RDF approach to assign a class to a resource (Bri, 2004). For example, the second triple in Figure 2 assigns the entity `DBpedia:David_Beckham` to the class `DBpedia:FootballPlayer`.

We look up the `rdf:type` of the entities that match the queries in the log file and count how often each type occurs. When multiple entities match a query, all entities are used. In this case, each matching entity is counted as $\frac{1}{n}$, where n is the number of matching entities.

Some entities have multiple types. For instance, `DBpedia:David_Beckham` is of type `DBpedia:FootballPlayer` as well as type `DBpedia:Person`. RDFS provides a property that defines a hierarchy of classes: `rdfs:subClassOf` (Bri, 2004). Using this property we find that `DBpedia:FootballPlayer` is a subclass of `DBpedia:Person`. We analyze the logs at two different levels: the lowest level (the smallest subclasses) and the highest level (the largest superclasses).
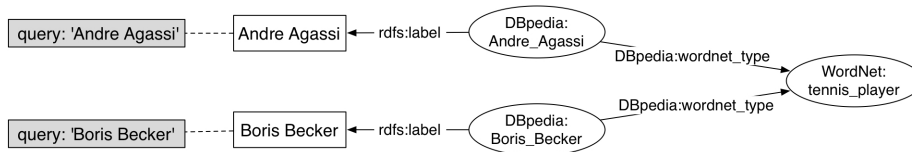
13

Figure 4: Example application of our method for finding semantic relations between queries: a relation between queries `Andre Agassi` and `Boris Becker` is that they both match DBpedia entities that are of WordNet type `tennis_player`.

Appendix B contains a complete description of our method for identifying query types.

3.4 Identifying semantic query modification patterns

To determine how users of a search engine modify their queries, we collect all pairs of queries that are consecutively submitted by the same user. For each pair of consecutive queries we determine the semantic relation between the queries, as illustrated in Figure 4. A graph search algorithm is used for traversing links in the linked data to find the shortest series of links that connects the entities matching the two queries (their relation).

In the next step we abstract away from relations between specific instances and infer *modification patterns* by removing the instances and keeping just the links. For instance, we may find that the relation from query `David Beckham` to query `Joe Cole` is that both refer to players in the English national football team:

> David Beckham –DBpedia:nationalteam→ England_natio-
> nal_football_team ←DBpedia:nationalteam–Joe Cole

The arrows denote the directions of the predicates. This relation is abstracted to the modification pattern:

> Q1 –DBpedia:nationalteam→ X ←DBpedia:nationalteam– Q2

To determine the importance of the patterns that are found, we count how often each pattern occurs between queries in the search log. The importance of a pattern is expressed by its *support* and *confidence*. Intuitively, the support value indicates how often a pattern occurs and the confidence corresponds to the likelihood that the pattern explains the observed modification well.

The support of a pattern p is defined as the relative number of query pairs matching p:

$$support_p = \frac{freq_p}{\sum_{p' \ in \ patterns} freq'_p}$$

Here $freq_p$ is the number of times pattern p occurs between consecutive queries in the search log and *patterns* is the set of all occurring patterns. To calculate the confidence of a pattern we first compute its *expected frequency*. Hereto,

we repeatedly sample at random two queries, each from a different session, and construct an artificial query pair, for which we determine the modification pattern. Patterns that occur often between random pairs are a priori more likely to occur and thus have a high expected frequency:

$$support\_random_p = \frac{freq\_random_p}{\sum_{p' \ in \ patterns} freq\_random_{p'}}$$

Here $support\_random_p$ is the expected frequency of pattern p and $freq\_random_p$ is the number of times pattern p occurs between sampled query pairs from different sessions. The confidence measures how much higher the support is than expected:

$$confidence_p = \frac{support_p}{support_p + support\_random_p}$$

A confidence value of 0.5 means that a pattern is totally unreliable, as it occurs equally often between consecutive queries as between pairs of randomly selected queries. A confidence of 1 means that the pattern has not been observed between any of the artificially constructed query pairs, which is a strong indication that the pattern results from the search strategy of the users and is not accidental. See Appendix C for the precise algorithm to compute support and confidence values.

In the last step of our method, we apply an iterative process to filter out patterns with low confidence and improve the accuracy of the relations that we find (see Appendix C).

3.5 Accuracy of the method

Assessing the accuracy of the proposed method for the automatic identification of semantic relations between queries is not trivial. The method consists of three steps: match the query to linked data entities, determine their types, and identify the modification patterns. Determining the appropriateness of a match between a query and a linked data URI can be difficult when the query is examined in isolation; for example, given a query corresponding to a person's last name, the URI matcher identifies a long list of people with the same last name - how to decide the 'correct' one without access to the individual who posed the query? Also, an attempt to measure the accuracy of the second step of the proposed approach, finding the query type, would actually boil down to assessing the quality of the linked open data used in the analysis. Given these considerations, we have not carried out a per-query evaluation for the first two steps of our approach, but let the outcomes of the analysis of the search logs, presented in Sections 4 to 6, speak for themselves[1].

Instead, we decided to investigate the accuracy of the end-to-end results only, by performing a (small-scale) user study to evaluate the modification patterns identified by the system. First, we sampled at random 100 pairs of consecutive

---

[1]Notice that Section 5 does compare the query types we identified with those found by an alternative automatic method.

queries from the search logs used in this study (see Section 4). Sampling has been limited to those query pairs where at most 10 modification patterns are identified (this is the case for the majority of query pairs that could be considered, and keeps the next steps workable). For this sample pool of query pairs, we ran the algorithm, and proceeded as follows to check the usefulness of the relations identified automatically by comparing them to a human standard.

Four people recruited from our research group (the 'judges') were asked to specify the 'most prominent relation' between query pairs. Each judge specified relations for 25 query pairs, so that in total we had relations for all 100 pairs. The judges were instructed to specify for example for query pair `bush, obama` that 'both are or were presidents of the US', while for `barrack obama, michelle obama` to state 'the second is the wife of the first' or just 'spouses'. We also included a purely term-based example in the instructions, telling judges to represent the relation between `michelle obarma, michelle obama` as 'same person, spelling corrected'. Because it is unlikely that anyone would know all the relations between the sampled query pairs, we instructed the judges to search the web for information on the entities found in the queries, and take ample time before resorting to a conclusion that 'no relation' exists. We allowed the judges to conclude that multiple relations are equally prominent, for example, to write 'sisters and famous tennis players' for the query pair `serena williams, vanessa williams`. We allowed a relation of 'queries unclear' for cases where the query would be too ambiguous, or incomprehensible (e.g., a query like `xgggeren1`). In the same way, a fifth judge independently created a second set of relations for all 100 query pairs to test the feasibility of the task and, to some extent, to check whether a judge would have spent sufficient effort in finding prominent relations. Finding the prominent relations for all sampled query pairs took between two and three hours.

Next, the three authors (the 'raters') individually rated each of the (at most 10) relations identified by the system, taking the viewpoint that the relations given by the human judges are correct; we repeated the rating for the two ground-truth relations found according to the human judging process just described (consisting of 4 times 25 and 1 time 100 judgements). Each of the raters classified the system's choice as either incorrect (0) (i.e., different from the judge's relation), approximately correct but more precise or less precise than the judge's relation (1), or correct (2) (i.e., the same as the judge's relation). We then computed agreement between judges and raters, where the 'no relation' pairs were treated as incorrect and the single pair classified as 'queries unclear' was disregarded.

Agreement between the two judgements over all separate assessments resulted in a Cohen's kappa of 0.75, while agreement between the three raters corresponds to a Fleiss-kappa of 0.69. We thus conclude that agreement between both judgements and raters is good. We mapped the ratings obtained into a range of $[0, 1]$, where we considered three mappings: *strict* maps correct to 1, and all other ratings to 0; *medium* extends the mapping with a score of 0.5 for the approximate correct ratings; and, *lenient* maps both correct and approximately correct to a value of 1. Computing the average of the mapped

rating per rater for each query pair, and averaging over all query pairs, results in a lenient score of 0.61 (and 0.33 for strict, 0.47 for medium).

Summarizing, the method is only moderately successful in identifying the most prominent relations for a given query pair. Nevertheless, as we will find in Section 6, the identified relations still provide good insights into the general user behavior. The evaluation of the method does also give interesting insights into the use of linked open data for this purpose. First of all, the algorithm often identifies relations between people based on 'uninteresting chance events'; for example, when human judges classify two queries as being related because both queries refer to 'British fashion models', the automatic method concludes that both queries refer to people born in London; similar cases occur for place of death. Also, DBpedia contains some inaccurate facts such as towns having nationalities, or people already divorced still labelled as spouses. Another interesting example is the query pair `angelina jolie, victoria beckham`, where the system identifies a relation of only limited value: that the spouse of the one person has the same profession as the other person; the 'true' relation identified by both human judges is of course that both are celebrities. Finally, a query pair `prince, queen` was classified as a relationship between royalty by one human judge, and as musicians by the other judge, which shows the inherent ambiguity of some of the query pairs. The system only found the former, and in the overall judgement therefore is rewarded only once.

4 Image search log data

The search log data in this study were collected over a 10-month period (October 2008 – July 2009) by the commercial picture portal of a European news agency that offers access to millions of photographic images to professional users, such as journalists, public relations agencies, and editorial staff in newspapers and magazines. The search actions that were recorded were 1) query submissions, 2) result viewings corresponding to clicks on selected images, and 3) image purchases corresponding to downloads of selected images. Each action was logged with a timestamp, session ID, the user's ID (in the case of users logged in to their account), and the submitted query. Click and download actions also logged the ID of the selected image. Downloading can only take place for logged in users.

Following Jansen (2006), we describe the log data in terms of *sessions*, *queries* and *terms*. We define a session as a series of consecutive search actions by a single user. Session boundaries are identified when the period of inactivity between two successive actions exceeds a 15-minute timeout. Queries are defined as the complete strings that are submitted in the search box. Parts of the queries separated by whitespace are called terms.

The collected search log data were preprocessed as follows in preparation for their analysis. First of all, the text of the submitted queries was 'lightly' normalized by converting it to lower case and removing punctuation, quotes, special characters, URLs, and names of major photo agencies. Furthermore,

empty queries and queries consisting only of whitespace characters were removed, while consecutive identical queries submitted in the same session were conflated into one.

Table 3 presents a summary of the statistics of the image search log data studied in this work[2]. After preprocessing, the search logs contain 1,094,620 queries, 21.2% of which are unique, part of 520,507 sessions. Out of these, 360,411 queries issued in 109,194 sessions correspond to interactions of logged in users; such interactions are performed by professional users who have an account in the news agency. Compared to previous image search log analyses, in particular those of professional users, the current study considers a much larger dataset (see Table 1).

Table 3: Summary statistics of the image search log data (Oct 2008 – Jul 2009).

Number of submitted queries	1,094,620
Number of unique queries	232,182
Percentage of queries that are unique	21.2%
Number of sessions	520,507
Number of queries per session	
Mean	2.1
Median	1
StdDev	3.13
Number of query pairs	574,113
Percentage of queries that are modified	52.4%
Number of terms per submitted query	
Mean	1.81
Median	2
StdDev	1.03
Number of terms per unique query	
Mean	2.25
Median	2
StdDev	1.73
Percentage of queries followed by click(s)	36%
Percentage of sessions containing click(s)	50%
Percentage of queries followed by download(s)	11%
Percentage of sessions containing download(s)	12%

The average number of queries per session is 2.1, which is lower than the average session length reported in previous web image search studies, where it ranges from 2.8 (Tjondronegoro et al., 2009) to 4.8 (Jansen et al., 2004) queries per session, and also in professional image search, where it corresponds to 3.3 queries per session (Jörgensen and Jörgensen, 2005) (see Section 2.1). For logged in users, the average session length is higher (3.3 queries per session), which is in line with previous findings. Users who are not logged seem to engage in

---

[2]It should be noted that the image search log data used in the present study are different to the datasets used in our previous work (Hollink et al., 2010a,b).

less complex interactions, also indicated by the high number of sessions (63.3%) comprising a single query (see Figure 5). One possible explanation is that some of the professional users begin their information searching without logging in and then after the first query, or the first couple of queries, they do log in and continue their search interactions. Unfortunately, this behavior cannot be detected by analyzing the search logs since as soon as users log in, they are assigned a different session ID and any link to their previous actions is lost.
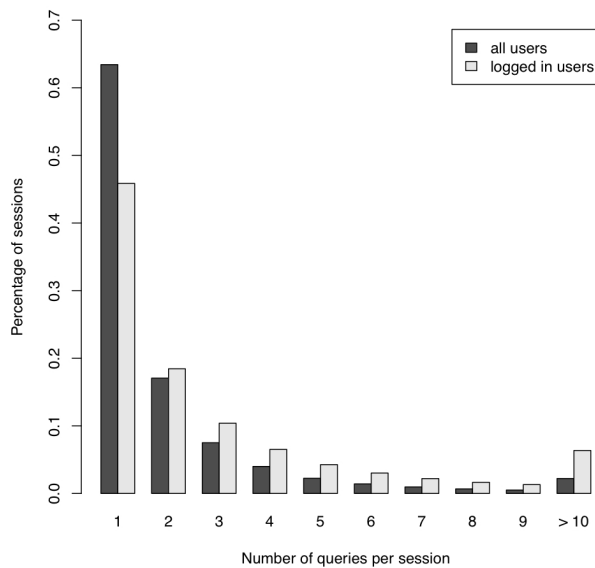

Figure 5: Relative frequency of session lengths for all users and for users who have logged in.

Furthermore, the distribution of queries per session indicates that 52.4% of the submitted queries are modified into another query. This results in 574,133 query pairs, i.e., pairs of consecutive queries submitted by the same user in the context of a single session; this figure reaches 69.7% for the case of users who are logged in. There is very little comparable research regarding the percentage of modified queries in web image search. In the context of professional image search, Jörgensen and Jörgensen (2005) report that 61.7% of queries are modified, which is in line with our findings. Query modifications are analyzed in detail in Section 6.

The submitted queries consist of approximately 1.8 terms, whereas if we consider only the unique queries, the average number of terms increases to around 2.2. This is lower than the number of terms per query reported for web image search (Jansen et al., 2004, Tjondronegoro et al., 2009, Goodrum and Spink, 2001, Jansen et al., 2000a), but it is close to the average query length

19

observed in the context of journalist image search (Jörgensen and Jörgensen, 2005, Westman and Oittinen, 2006). A look at the most frequently occurring queries in Table 4 reveals that they are indeed short and mostly concerned with named entities, in particular specific persons. In conjunction with the findings of previous research in professional image search, there appears to be a fairly strong trend in submitting short queries in this context. Figure 6 illustrates the frequency distribution of the queries which, as expected, follows a power law with a long tail of queries submitted only a handful of times.

Table 4: The twenty most frequently occurring queries and the number of times they occur.

	# Occ.	Query		# Occ.	Query
1.	11,844	cristiano ronaldo	11.	3,130	prince
2.	7,152	fernando torres	12.	3,115	karim benzema
3.	6,445	letizia	13.	3,098	princess letizia
4.	6,197	princess mary	14.	3,089	spain
5.	5,866	princess	15.	3,077	bruel
6.	4,996	russell brand	16.	2,999	monaco
7.	4,194	beckham	17.	2,917	princess mathilde
8.	3,382	sergio ramos	18.	2,870	belgium
9.	3,250	carla bruni	19.	2,798	mathilde
10.	3,163	brad pitt	20.	2,742	nicolas sarkozy

The image search results are displayed as image thumbnails, enabling users to make an informed decision before clicking on a particular image so as to preview it in bigger size and view more of its accompanying metadata. Table 3 indicates that about a third of the submitted queries are followed by at least one click and about half the sessions have at least one result followed up. This number of sessions containing at least one click is in line with the 56% previously reported (Tjondronegoro et al., 2009), whereas the number of queries followed through is higher than the 26% previously observed in professional image search (Jörgensen and Jörgensen, 2005). This could be due to advancements in the effectiveness of the underlying retrieval technology in recent years that allow users to engage in more successful searches. This may also explain the higher number of queries followed by an image download (11%) and of sessions containing at least one download action (12%) compared to the previously reported 7.2% of searches that resulted in downloads in a similar professional setting (Jörgensen and Jörgensen, 2005). This rise is further evident if we only consider the logged in users, since they are the only ones that can purchase an image by downloading it, for whom these figures correspond to 33% of their queries and 60% of their sessions.

In conclusion, users who are searching for images professionally tend to submit short queries. On average their sessions are not long, but there is a significant number of longer sessions and of queries that are iteratively modified. The semantic analysis performed in the two following sections aims to shed more
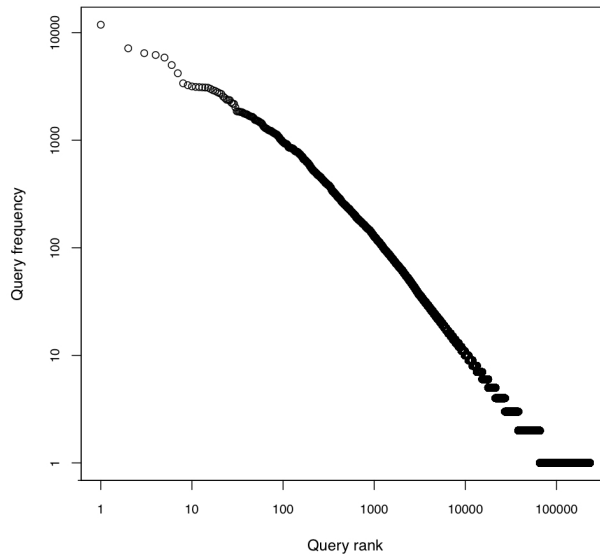
Figure 6: Rank-frequency distribution of the submitted queries ordered by rank using a log-log transformation.

light on the types of the submitted queries as well as the relations between consecutive queries.

5   Query type analysis

In the following we analyze the types of the news agency queries with the method described in Section 3.3 and show how this method allows one to easily switch between different levels of types. At the highest level, a distinction is made between *conceptual* and *specific* queries (e.g., Westman and Oittinen, 2006, Hollink et al., 2004). Conceptual queries (also called generic queries) describe general objects, such as `wall` or `fishing`, while specific queries ask for specific instances such as `the Berlin Wall`. Specific queries are analyzed by looking at the frequency of semantic types, such as persons and places, and of subtypes of these types, such as soccer players. For conceptual queries, we determine how often queries consist of the various lexical categories (nouns, adjectives, etc.) and we classify the nouns according to the top-level categories of WordNet (Fellbaum, 1998), which include categories such as `objects, events` and `phenomena`.

Section 5.1 discusses the types that were found when analyzing all queries. Section 5.2 compares our findings to the results of an existing method for determining query types. In Section 5.3 we compare the types of queries that were

successful to the types of queries that were not successful.

5.1 Types of all queries

For 79% of the 1,094,620 queries a matching URI was found. For 68% of the queries with a URI the type could be determined. Table 5 shows the ten most frequently occurring query types when the types are analyzed at the highest level[3] (see Section 3.3). The table contains various types that we would view as similar, such as `DBpedia:Place`, `tgn:AdministrativePlace`, and `tgn:PhysicalPlace` (all place names). Our algorithm does not recognize these types as being similar as there is currently no link between the classes in our linked data. When such links are added (which is the intention in the linked data community), equivalent classes can be merged, leading to a more comprehensive view on the query type frequencies.

Table 5: The ten most frequently occurring query types found when analyzing query types at the highest level and the relative frequencies of the types for all queries (all), for successful queries (suc.), and for unsuccessful queries (unsuc.).

	all	suc.	unsuc.	type	example query
1.	0.440	0.424	0.433	`DBpedia:Person`	`cristiano ronaldo`
2.	0.103	0.098	0.108	`tgn:AdministrativePlace`	`berlin`
3.	0.053	0.054	0.058	`WordNet:NounSynset`	`princess`
4.	0.024	0.025	0.024	`DBpedia:Place`	`monaco`
5.	0.016	0.018	0.015	`DBpedia:Organisation`	`real madrid`
6.	0.013	0.012	0.015	`aat:Concept`	`shoe`
7.	0.007	0.006	0.007	`DBpedia:Work`	`out of africa`
8.	0.006	0.006	0.006	`tgn:PhysicalPlace`	`santa barbara`
9.	0.004	0.004	0.005	`WordNet:VerbSynset`	`swim`
10.	0.003	0.003	0.002	`DBpedia:Species`	`snailfish`

In line with Jörgensen and Jörgensen (2005), Westman and Oittinen (2006), and Markkula and Sormunen (2000), we found that conceptual searches were less common than searches for specific instances. Table 5 contains three conceptual types: `WordNet:NounSynset`, `WordNet:VerbSynset` (nouns and verbs in WordNet), and `aat:Concept` (concepts from the Getty AAT). Examples of specific types are `DBpedia:Person` (person names) and `tgn:AdministrativePlace` (place names).

For specific queries, our method determines the semantic type of the queries. As shown in the table, by far the most common type of specific queries is `DBpedia:Person`. Other frequent types are places and organizations. Large numbers of queries for persons are also found in Tjondronegoro et al. (2009), Jörgensen and Jörgensen (2005), Westman and Oittinen (2006), and Markkula

---

[3]Initially, our analysis gave `DBpedia:Resource` as the most common type, because all DBpedia entities are both of this type and one or more other types. As having `DBpedia:Resource` as type is meaningless, we removed it from the analysis.

Table 6: The ten most frequently occurring subtypes of `DBpedia:Person` found when analyzing query types at the lowest level and the relative frequencies of the types. Here `DBpedia:Person` is the type of person queries for which no subtype is known.

	freq.	type	example query
1.	0.123	DBpedia:FootballPlayer	cristiano ronaldo
2.	0.100	DBpedia:Actor	brad pitt
3.	0.083	DBpedia:MusicalArtist	lara fabian
4.	0.029	DBpedia:TennisPlayer	rafael nadal
5.	0.022	DBpedia:Person	sarah cawood
6.	0.019	DBpedia:Comedian	russell brand
7.	0.017	DBpedia:OfficeHolder	nicolas sarkozy
8.	0.012	DBpedia:Model	claudia schiffer
9.	0.012	DBpedia:Cyclist	tom boonen
10.	0.004	DBpedia:PokerPlayer	michael greco

and Sormunen (2000). Apparently, access to image repositories is often motivated by a need to depict a person.

The semantic analysis method can also describe query types at lower levels, which allows one to look deeper into the use of particular semantic query types. Table 6 shows the most commonly searched subtypes of persons. As shown, many people searched for sports people, which is indeed a topic on which the news agency repository contains a large number of photos. Other common person types are actors, musicians, and comedians. The news agency can use this information when deciding on future additions to the repository.

Conceptual queries can be categorized according to their lexical categories (part-of-speech tags). If we look at the various query types from WordNet (`WordNet:NounSynset`, `WordNet:VerbSynset`, etc.) we find that nouns are by far the most common lexical category (52%), followed by verbs (26%), adjectives (20%) and adverbs (1%). The dominance of noun concepts is also found in Jörgensen and Jörgensen (2005), but they find more adjectives than verbs. Manual inspection of the identified verbs showed that our algorithm classified many queries as both nouns and verbs (e.g., `milk, miss`). Most likely, such queries are meant as nouns rather than verbs (Bozzon et al., 2007), so the verb count should be lower indeed.

We semantically analyzed the queries with type `WordNet:NounSynset` by following for these queries the `WordNet:hyponymOf` relation instead of the `rdf:type` relation. In this way these queries are classified into the top-level classes of the WordNet noun hierarchy (see Fellbaum (1998)). For instance, the query `horse race` is now classified as an `event`. The frequency of the ten top-level noun categories are given in Table 7. The most common category is `entity`, which includes physical objects and life forms. Examples of queries categorized as entities are `telephone booth` and `nanny`. Groups, acts, and abstractions are also seen quite regularly. The other categories appear much less frequently.

Table 7: The nine top-level categories of `WordNet:nouns` and the relative frequencies of the categories.

	freq.	type	example query
1.	0.0332	entity	princess
2.	0.0022	group	french foreign legion
3.	0.0022	act	cycling
4.	0.0020	abstraction	solidarity
5.	0.0010	psychological feature	haute couture
6.	0.0009	state	flu
7.	0.0006	event	tour de france
8.	0.0003	phenomenon	global warming
9.	0.0001	possession	christmas presents

5.2 Comparison to OpenCalais

For comparison, we have followed Tjondronegoro et al. (2009) in using Open-Calais to classify queries (see Section 2.1). This Web Service is really intended for the annotation of natural language text such as newspaper articles, and is not meant to be applied to the short queries found in search logs. Yet, when submitting the search logs in batches of 500 queries at a time (one query per line), only few batches cannot be analyzed (these raise a language detection error).

The ten most frequent types identified by OpenCalais are summarized in Table 8. Other prominent types are product, region, facility, and event. Approximately three quarters of the queries remain unclassified. Because the Web Service may assign types to part of the query text only, and also identify multiple types per query, we give absolute numbers rather than percentages and these cannot be compared directly with the results of our own type analysis. However, the main findings are consistent with ours, in the significant number of queries targeting people and locations. When comparing our method to these results, the type analysis based on linked data that we developed can classify types at different levels, such as person versus musician. Types returned by OpenCalais such as industry and technology correspond really to domains. We conclude that both approaches are complementary, having their own strengths and weaknesses, but, as far as they are comparable, lead to consistent results.

5.3 Types of successful and unsuccessful queries

We now move on to analyzing differences in query types between successful and unsuccessful queries. A successful query at least partially answers the user's information need. Therefore, similar to Huang and Efthimiadis (2009), we define a successful query as a query that is followed by at least one click on a search result. Unsuccessful queries do not result in a click and are followed directly by another query or end the search session. According to this definition, in our data 389,193 queries (36%) were successful and 687,869 queries (64%) were

Table 8: The ten most frequent query types according to OpenCalais.

	count	type	example query
1.	174,694	Person	`brad pitt`
2.	68,192	Position	`real madrid goalkeeper`
3.	4,583	IndustryTerm	`belgium bank`
4.	3,377	SportsEvent	`tour de france`
5.	2,308	SportsGame	`soccer`
6.	1,683	Organization	`whitehouse`
7.	1,379	Technology	`gps`
8.	1,244	City	`brussels`
9.	1,071	Holiday	`christmas`
10.	700	Country	`belgium`

unsuccessful (see Section 4).

Alternatively, we can define success by a download instead of a click. Downloading (and thus paying for) an image is a stronger indication that the user is satisfied with the search result than just clicking on the image. However, because the results for both definitions of success turned out to be very similar and most other studies only have access to click information, we present the results of the click analysis only.

The second and third columns of Table 5 show the frequency of the ten query types for successful and unsuccessful queries. Conceptual queries are not very successful: types from WordNet and the AAT occur more often among the unsuccessful than among the successful queries. This is consistent with our expectations. Conceptual queries are considered to be difficult, as a result of differences between the terminology used in queries and image annotations (Markkula and Sormunen, 2000). Most concepts can be described by various synonyms, so that conceptual queries often contain different terms than the annotations of the corresponding images. Moreover, the annotations often contain descriptions of specific entities, such as 'Empire State Building' rather than general concepts such as 'skyscraper'.

Finally, we observe that successful queries could be mapped more often to URIs (79%) that unsuccessful queries (76%). This indicates that more common topics, which have a larger probability of occurring in the linked data, are on average more successful.

In conclusion, our results showed that the semantic analysis can effectively determine semantic query types at various levels of specificity. The analysis confirmed the findings of previous studies that conceptual queries are used less frequently than queries for specific entities and that persons are the most common type of specific entities. We also showed that specific queries were more effective than conceptual queries. In addition, the presented method enabled us to break down query types into subtypes, showing, for example, that in our domain users often search for football players and actors. For conceptual queries,

we found that users mainly use nouns, in particular entities.

6 Query modification analysis

With the method for finding semantic query modification patterns presented in Section 3.4, we analyzed the pairs of queries occurring consecutively in the sessions found in the search log. The results of this analysis are compared to the results of a term-based analysis. We start with a discussion on query modifications found when analyzing all queries and then zoom in on differences between modifications made after successful and unsuccessful queries.

6.1 Modifications of all queries

For 24% of the 574,113 query pairs a relation was found. The ten patterns with the highest support are shown in Table 9. These patterns all have confidence values that are close to 1, which means that it is unlikely that they occurred as chance events instead of representing search strategies of the users.

The most common pattern was the identity relation ([ ]): two different queries that are mapped to the same entity, usually variant names for the same entity, such as `Gent` and `Gand` (the Dutch and French name of a Belgian city). These do not include identical queries, as they are conflated during preprocessing (see Section 4). Patterns 2, 6, and 10 indicate that many users searched first on the name of a person and then on the name of his or her spouse or partner. Patterns 3, 9, and 4 tell us that many users successively searched for two people with a common property: respectively, people who play for the same national team, who belong to the same sports club, and who star in the same movie. Patterns 5 and 8 both say that users searched on two entities with the same type, such as tennis players or townships. Pattern 7 uses the `AAT:distinguished_from` relation from the Getty Art and Architecture Thesaurus which links closely related terms, such as `prince` and `princess`.

Inspection of the modification patterns reveals two important classes of modifications (see Table 10). 19% of the identified relations are what we refer to as *sibling relations*: relations of the form `Q1–R→X←R–Q2` or `Q1←R–X–R→Q2`. Examples include patterns 3, 4, 5, 8, and 9 in Table 9. This shows that many people searched for two entities with some common property, such as two actors starring in the same movie or two hyponyms of a WordNet concept. This finding is in line with the work of Rieh and Xie (2006), who found through manual classification of query modifications that sibling relations (referred to as 'parallel movements' by them) were the most common class of modifications. However, their analysis did not identify what kind of siblings were used (actors, or soccer players, or hyponyms, etc.). The second frequently occurring class of relations (10%) are *direct few-to-few relations*, which are defined as a relaxed version of one-to-one relations, where 'few' means on average less than 2. Examples of direct few-to-few relations are 'spouse' and 'has-capital'.

Table 9: The ten semantic modification patterns with the highest support with their support and confidence values.

	sup.	conf.	pattern
1.	0.025	0.96	[ ]
2.	0.017	1.00	Q1 —DBpedia:spouse→ Q2
3.	0.014	0.96	Q1 —DBpedia:nationalteam→ X ←DBpedia:nationalteam— Q2
4.	0.013	0.98	Q1 ←DBpedia:starring— X —DBpedia:starring→ Q2
5.	0.011	0.91	Q1 —rdf:type→ X ←rdf:type— Q2
6.	0.010	1.00	Q1 —DBpedia:partner→ Q2
7.	0.008	0.99	Q1 ←aat:distinguished_from— Q2
8.	0.008	0.93	Q1 —DBpedia:wordnet_type→ X ←DBpedia:wordnet_type— Q2
9.	0.005	0.95	Q1 —DBpedia:clubs→ X ←DBpedia:clubs— Q2
10.	0.005	0.98	Q1 ←DBpedia:spouse— Q2

Table 10: Relative frequency of classes of semantic query modification patterns (proportion of identified query pairs with identified relations) after all queries, after successful queries and after unsuccessful queries.

Modification type	all queries	successful queries	unsuccessful queries
sibling	0.193	0.228	0.178
few-to-few	0.105	0.069	0.112
other relations	0.702	0.703	0.710

We also analyze the query pairs with a term-based approach (see Section 2.2). After stemming the query terms, we determine whether, compared to the previous query, terms are added (specification), removed (generalization) or replaced (reformulation). In addition, we count how many times stemming made the query identical to the previous query (stem-identicals). Query pairs without overlapping terms are classified as 'undetermined'. The frequency of each type of term modification is shown in Table 11 and Figure 7.

25% of the modifications can be assigned to one of the four term-based classes, which is comparable to the result of term-based approaches in previous studies (e.g. 23% in Jones and Fain (2003), see Section 2.2). Reformulations occur most frequently: 1.5 times as often as specifications, which occur 1.9 times as often as generalizations. Stem-identical queries are the least frequent class. These findings closely match the findings of previous studies (Bruza and Dennis, 1997, Lau and Horvitz, 1999, Jörgensen and Jörgensen, 2005, Rieh and Xie, 2006, Whittle et al., 2007, Jansen et al., 2009, Boldi et al., 2009, Costa and Seco, 2008, Efthimiadis, 2000, Jones and Fain, 2003).

Comparing the results of term-based and semantic analysis, we see that the percentage of the modifications that can be assigned to one of the four term-based classes (25%, see Table 11) closely matches the percentage of cases for which a semantic modification could be found (24%). However, as shown in Figure 8, the two approaches classify different query pairs: the linked data

Table 11: Relative frequency of term-based query modification types after all queries, after successful queries and after unsuccessful queries.

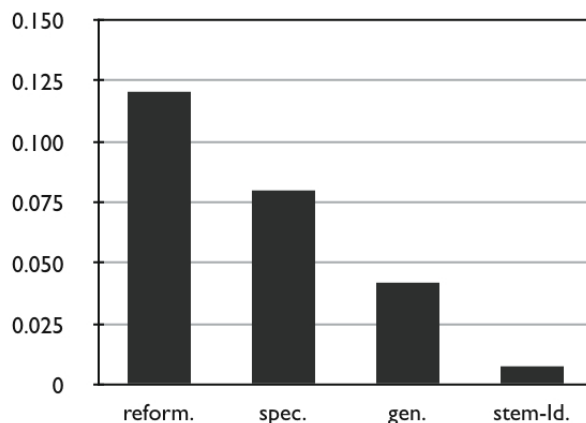| Modification type	all queries	successful queries	unsuccessful queries
specification	0.080	0.056	0.092
generalization	0.042	0.022	0.052
reformulation	0.121	0.102	0.136
stem-identical	0.008	0.005	0.010
undetermined	0.750	0.815	0.711

Figure 7: Relative frequency of term-based modification types.

approach found a relation for only 9% of the cases that were classified by the term-based approach. Conversely, the term-based approach found a class for only 9% of the cases for which a semantic relation was found, which shows that the two analyses are complementary. One reason for this effect is that the term-based approach works well for cases in which names of multiple entities are present in one query, such as 'Beckham Milan' and 'Beckham Madrid', while it cannot handle most pairs of queries consisting of single entities, such as the pair 'Andre Agassi' and 'Boris Becker'. The linked-data approach that we developed, on the other hand, can currently handle single entity queries, but not multiple entity queries. In the future, the linked-data method may be extended to handle queries with multiple entities. This requires queries to be split up in meaningful fragments and each fragment to be matched separately on a linked-data entity.

Reformulations are often related to sibling relations: for example, the modification from the query elm tree to the query oak tree is both a reformulation and a sibling relation. Nevertheless, a large part of the siblings found with the linked data approach cannot be recognized by looking at term-based reformulations: only 7% of the siblings were recognized as reformulation. Inspection of the results shows that many siblings consist of names of two persons, such as two
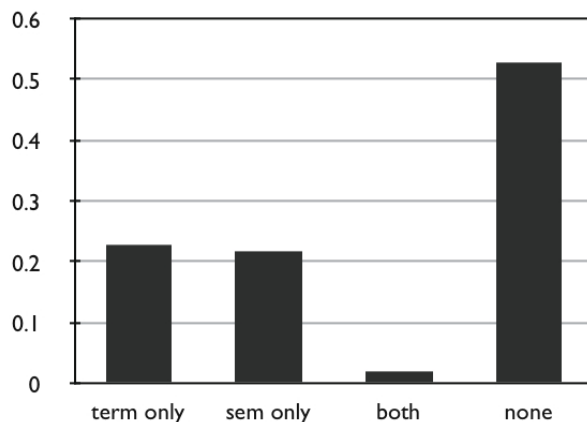
Figure 8: Overlap between the query pairs classified by the semantic and the term-based approach: the proportion of queries for which a relation is found only by the term-based approach, only by the semantic approach, by both approaches, and by none of the approaches.

players in the same national team. The names do not have any terms in common and thus are not classified as reformulations. In other words, the queries were semantically related, but not with respect to their terms. There is no corresponding term-based class for direct few-to-few relations. These relations are classified by the term-based approach as reformulations (2%), specifications (0.1%), generalizations (0.1%) and stem-identical (0.2%), but most often as undetermined (98%). This shows that there are types of modifications that are important for users, but that cannot be identified with a term-based analysis of query modifications.

6.2 Modifications of successful and unsuccessful queries

We analyzed differences in query modifications of successful and unsuccessful queries using the same definition of success as in Section 5.3. Like before, considering downloads instead of clicks led to similar results.

The frequency of the four term-based modification types after successful and unsuccessful queries are shown in Table 11. Successful queries are more often followed by a modification that is classified as undetermined (81%) than unsuccessful queries (71%). This could mean that after successful queries users more often switch to unrelated topics, probably because they were satisfied with the results they found for the first search topic. Unsuccessful queries, on the other hand, are more often followed by stem-identical queries (1.0% vs 0.5%), which indicates that after unsuccessful queries users more often make a second attempt to find the same information using an almost identical query. This finding is in line with the work of Huang and Efthimiadis (2009), who found some forms of term variations often occur when no clicks have been made (see

Section 2.2). Unsuccessful queries also lead to more specifications, generalizations, and reformulations than successful queries do. Most likely, this can be explained in the same way as the high number of stem-identical queries: users making a second attempt to find the same information. The effect was strongest for generalizations, which appear 2.3 times as often after unsuccessful queries than after successful queries. Apparently, generalizations are often used to expand the result set when no relevant results were found for the initial query. This contradicts the findings of Huang and Efthimiadis (2009): they found that generalization and reformulation occur relatively often after a click has been made.

We also identified which semantic modification patterns occur frequently after successful and unsuccessful queries. To this end we applied our method for finding semantic modification patterns to the set of query pairs for which the first query was successful and the set of query pairs of which the first query was unsuccessful. The ten patterns with the highest support after successful queries are shown in Table 12. For each pattern the figure also shows the support of the pattern after unsuccessful queries.

Table 12: The ten semantic modification patterns with the highest support for successful queries with their support values for successful queries (suc.) and unsuccessful queries (unsuc.).

	suc.	unsuc.	pattern
1.	0.019	0.012	`Q1 −DBpedia:nationalteam→ X ←DBpedia:nationalteam− Q2`
2.	0.015	0.030	`[ ]`
3.	0.013	0.010	`Q1 −rdf:type→ X ←rdf:type− Q2`
4.	0.011	0.013	`Q1 ←DBpedia:starring− X −DBpedia:starring→ Q2`
5.	0.011	0.018	`Q1 −DBpedia:spouse→ Q2`
6.	0.009	0.007	`Q1 −DBpedia:wordnet_type→ X ←DBpedia:wordnet_type− Q2`
7.	0.007	0.004	`Q1 −DBpedia:clubs→ X ←DBpedia:clubs− Q2`
8.	0.005	0.004	`Q1 −aat:distinguished_from→ Q2`
9.	0.005	0.004	`Q1 −DBpedia:occupation→ X ←DBpedia:occupation− Q2`
10.	0.005	0.005	`Q1 ←DBpedia:starring− X −DBpedia:distributor→ Y`
			`←DBpedia:distributor− Z −DBpedia:starring→ Q2`

Successful queries are more often than unsuccessful queries followed by a modification for which no semantic relation could be found: no relation was found for 58% of the modifications of successful queries for which entities in the linked data could be found, versus 54% for unsuccessful queries. In all likelihood, after successful queries users are more inclined to switch their search topic, as we found with the term-based analysis. Like stem-identicals, the identity relation ([ ]) occurred almost twice as often after unsuccessful queries than after successful queries (support 0.030 vs 0.015), probably due to users making a second attempt to find the same information.

As shown in Table 10, sibling relations were observed most frequently after successful queries (23% for successful queries versus 18% for unsuccessful

queries). Table 12 shows various examples of such relations: two entities of the same type (patterns 3 and 6), two players from the same national team (pattern 1) or club (pattern 7), and two people with the same occupation (pattern 9). An explanation for this phenomenon can be that in these cases the first query only covers part of the users' larger information need. Once this part is satisfied, the user moves on the next part. For example, a user who is looking for images of a soccer team in action may pursue this goal by consecutively querying on the names of various star players from the team.

Direct few-to-few relations occurred less often after successful queries (7%) than after unsuccessful queries (11%), which suggests that these modifications are used to find an image that shows the subject of the user's information need in a different context. For example, users looking for images of the private life of David Beckham but only finding soccer action photos may query for Victoria Beckham (his spouse) as a second attempt.

In conclusion, the semantic approach for analyzing query modifications forms a valuable addition to term-based approaches. With the semantic method we identified a number of modification patterns that are frequently employed by users of the search engine, but that do not become visible in term-based analyses. We found that many query modifications can be classified as sibling relations and direct few-to-few relations. Sibling relations are used most often after successful queries while few-to-few relations occur mainly after unsuccessful queries. As far as the results of semantic approach and the term-based approach are comparable, their findings are consistent: both approaches showed that users often try variant names for the same entity, especially when the first query was not successful.

7 Impact and applications

The results of the presented study have implications for the design of search engines and search support tools. In this section we discuss various applications. A summary of these applications and the findings that can be used for these applications is given in Table 13.

Several search tools make use of ontologies, thesauri or lexicons, for example to generate terms for query suggestions or automatic query expansion (Hollink et al., 2007) or to cluster search results (Hemayati et al., 2007, Ren et al., 2009). Semantic search log analysis enables us to discover which types of information users frequently search for and which relations between queries are frequently used. This information can be used to improve the match between the resources that are used and the information needs of the users. Through the picture portal of the news agency that we studied users most often search for persons, in particular football players, musical artists and actors. Another frequent category were geographical locations. General concepts, such as common nouns and verbs, occurred much less frequently. The relations between queries that were observed most frequently were variant names for the same entity, direct

Table 13: Summary of potential applications of the findings in this study.

Application	Findings
• Improving the match between resources used by search tools and users' information needs	Query types and relations that are frequently searched for
• Selection of concepts for image content analysis	Query types that are frequently searched for
• Purchasing and annotating images for an image collection	Types of queries that frequently do not lead to clicks
• Differentiating query suggestions between successful and unsuccessful queries	Differences between modification patterns after successful and unsuccessful queries
• Increasing recall of image search	Frequently used relations between queries
• Providing search feedback	Frequently used properties shared by sequences of queries
• Improving session boundary detection	Relations between consecutive queries that do not have terms in common

relations between persons, and, category membership. The observed types of entities should guide the selection or construction of ontologies or lexicons for a search engine. Currently, the most frequently used resource is WordNet (e.g., Hollink et al., 2007). Our findings suggest that this is probably not the best choice, as WordNet contains mainly general terms, very few person and place names and no relations between persons; exploring Wikipedia (or DBpedia) seems a promising alternative resource for this purpose.

Our findings also have implications for automatic image annotation through content analysis. A crucial step in content analysis is the selection of concepts for which a classifier is built (Lin and Hauptmann, 2006). The frequent query types can help to choose those concepts that match the user's query. Our results indicate also that face recognition could be a valuable type of content analysis for the automatic annotation of journalistic images.

The identified frequent types of queries can also guide the focus of the news agency when purchasing new images to be added to their collection. Images that match the users' information needs have a higher probability of being sold and thus are a more valuable addition to the collection. This strategy is complementary to the strategy of acquiring images from categories with high sales numbers. In contrast to the latter strategy, our methodology also reveals query types that do not currently result in a click on an image. Knowing about these query types is useful for managing the image collection. If the collection does not contain images matching such queries, adding such images will probably increase sales. If relevant images are in the collection, but they are nonetheless not found, apparently the image annotations do not match the users queries and need improvement.

Current search assistance tools, such as query suggestion generators, do not distinguish between successful queries (queries that have led to a click on a search result) and unsuccessful queries. However, our experiments show that searchers behave quite differently after successful and unsuccessful queries. Taking these differences into account can potentially enhance the effectiveness of search support. For example, our results indicate that offering variant names for the entity on which the user has searched is more important after unsuc-

cessful queries, while showing related entities is more important after successful queries.

The identified modification patterns can be applied to increase recall when no or few results are found. In these situations the search engine can return results matching entities that are related to the user's query. The frequent modification patterns indicate which relations are most likely to be of use. Using these relations also enables the system to explain the relation between the query and the returned results.

Our results can also be used for providing search feedback. We found that users often list a number of entities with a common property, such as players from the same soccer team. By matching queries to common modification patterns, we can recognize such query sequences at an early stage. If the image collection does not contain any images of entities with the property that was searched for, the system can inform the user early on that the current line of search is not going to be successful, saving the user time and frustration. If the collection does contain such images, the system can show the names of the entities as search suggestions, allowing the user to recognize rather than recall the entities, which has shown to be cognitively less demanding (Molich and Nielsen, 1990, Nielsen, 1994).

The approach that we used to determine relations between queries can be applied to detect session boundaries. Existing session boundary detectors are usually based on the time between requests or the overlap in terms between queries (e.g., He et al., 2002, Özmutlu, 2006). We showed that the latter approach is not sufficient as many related queries do not have terms in common. Complementing term overlap with semantic relatedness of queries can potentially improve the accuracy of the detected boundaries. Similarly, semantic relations may be applied to recognize multitasking behavior: users who are searching for two or more unrelated topics at the same time.

The presented findings are restricted to image search in a journalistic context. Further research is needed to determine to what extent the identified patterns are generalizable to other contexts. The methodology, on the other hand, is generally applicable to search logs. In other contexts the semantic analysis may yield different search patterns, but these patterns can be applied in the same manner to optimize search support in these contexts.

8   Conclusions

We presented a methodology to analyze search logs on a semantic basis using information from linked data sources. Our experiments show that this methodology can effectively determine which types of queries are frequently used and which relations exist between pairs of queries that are consecutive entered in a search session. The presented method sufficed to describe query types at a high level as well as to discover specific subtypes. The semantic approach proved complementary to existing term-based approaches as the two types of approaches reveal different patterns in the users' information needs and search

strategies.

We used the semantic analysis method as well as a statistical analysis to study the search behavior of users who professionally searched in an image repository. The analyses showed that these users tend to use very short queries (1.8 terms on average). Queries often consist of person names (44% of the queries) and other named entities. Queries consisting of general concepts were used less often. The general concepts that were used were most often nouns, in particular entities. Queries for general concepts proved more difficult than queries for named entities: they ended less often in a click on an image.

A term-based analysis of query modifications confirmed findings of earlier studies: the most frequently observed modification is replacing terms in a query by other terms, followed by adding terms to a query, removing terms from a query, and, least frequently, using lexical variations of terms. The semantic query modification analysis gave new insights. It showed that users frequently tried a number of queries consisting of variant names for the same entity (5% of the identified modifications), especially after the first query did not lead to a click on a search result. Users also often listed entities with a common property, such as actors starring in the same movie or concepts with a common hypernym (19%). This pattern was observed most often after a click had been made, suggesting that users often spread their information need over several queries. Finally, 10% of the modifications could be described by domain-specific relations between named entities, such as `spouse-of`, a pattern that was seen mainly when no click was made.

We discussed implications of our findings for various applications. We explained how the outcomes of our analysis can be used to improve the effectiveness of various types of search support and for the detection of session boundaries. In addition, we discussed the impact of our results on the management and annotation of journalistic image collections.

In the future we are planning to broaden the scope of our research to the analysis of image search in other domains and other types of users. Also, we will look at search strategies that go beyond query pairs and describe search behavior during whole search sessions.

9   Acknowledgements

A   Algorithm for mapping queries to linked data entities

Input: *sessions*: list of search sessions

triples: all RDF triples in the used linked data sources
Output: *query−entities*: list of queries and matching entities

query−entities = ∅
for each session s in *sessions*
    Q_s = list of queries in session s ordered by time
    for each query q in Q_s
        $matching\_entities_q$ = ∅
        for each triple $< subject, \texttt{rdfs:label}, object >$ in *triples*
            $label$ = lowercase(*object*)
            if $label = q$
                add *subject* to $matching\_entities_q$
        if $matching\_entities_q$ = ∅
            q' = remove_accents(q)
            q_{stem} = porter_stem(q')
            for each triple $< subject, \texttt{rdfs:label}, object >$ in *triples*
                $label_{stem}$ = porter_stem(remove_accents(lowercase(*object*)))
                if forall terms t in q_{stem} it holds that t is in $label_{stem}$
                    add *subject* to $matching\_entities_q$
    add $matching\_entities_q$ to *query−entities*

B   Algorithm for identifying query types

Input: *query−entities*: list of queries and matching entities
              sessions: list of search sessions
              triples: all RDF triples in the used linked data sources
              mode: most_specific_types or most_general_types
Output: *query−type−frequencies*: list of query types and their relative frequencies

for each session s in *sessions*
    Q_s = list of queries in session s ordered by time
    n_q = number of queries in Q_s
    for each query q in Q_s
        $entities$ = set of entities matching q according to *query−entities*
        n_e = number of entities in *entities*
        for each entity e in *entities*
            $types_e$ = ∅
            for each triple $< e, \texttt{rdf:type}, object >$ in *triples*
                add *object* to $types_e$
            for each type t in $types_e$
                if $mode$ = lowest_level_types
                    if there is a t' in $types_e$ such that $< t', \texttt{rdfs:subClassOf}, t >$ in *triples*
                        remove t from $types_e$
                if $mode$ = highest_level_types

$\quad\quad\quad\quad\quad$ **if** there is a t' in $types_e$ such that $< t, \texttt{rdfs:subClassOf}, t' >$ in $triples$

$\quad\quad\quad\quad\quad\quad$ remove t from $types_e$

$\quad\quad\quad\quad$ n_t = number of types in $types_e$

$\quad\quad\quad\quad$ **for each** type t in $types_e$

$\quad\quad\quad\quad\quad\quad$ $frequency_t \mathrel{+}= 1/(n_q \cdot n_e \cdot n_t)$

all_types = set of types t for which $frequency_t$ is defined

for each type t in all_types

$\quad\quad$ $normfrequency_t = frequency_t / \sum_{t' \ in \ all\_types} frequency_{t'}$

$\quad\quad$ add $normfrequency_t$ to $query{-}type{-}frequencies$

C $\quad$ Algorithm for identifying query modification patterns

Input:	$query{-}entities$: list of queries and matching entities
	$sessions$: list of search sessions
	$triples$: all RDF triples in the used linked data sources
	$support\_threshold$: minimum support (set at 0.0005)
	$confidence\_threshold$: maximum confidence (set at 0.66667)
Output:	$modification{-}pattern{-}support$: list of query modification patterns
	$\quad\quad$ and their relative frequencies
	$modification{-}pattern{-}confidence$: list of query modification patterns
	$\quad\quad$ and their confidence scores

$taboo\_patterns = \emptyset$ (list of modification patterns that are not allowed)

 # Step 1: sample random query pairs

$total\_number\_of\_pairs$ = total number of query pairs in $sessions$

$random\_query\_pairs = \emptyset$

$cur\_number\_of\_pairs = 0$

while $cur\_number\_of\_pairs < total\_number\_of\_pairs$

$\quad\quad$ s_1 = a random session from $sessions$

$\quad\quad$ s_2 = a random session from $sessions$ such that $s_1 \neq s_2$

$\quad\quad$ q_1 = a random query from s_1

$\quad\quad$ q_2 = a random query from s_2

$\quad\quad$ add $< q_1, q_2 >$ to $random\_query\_pairs$

$\quad\quad$ $cur\_number\_of\_pairs = cur\_number\_of\_pairs + 1$

 # Step 2: count modification patterns between consecutive query pairs

for each session s in $sessions$

$\quad\quad$ $pairs_s$ = set of consecutive query pairs in s

$\quad\quad$ n_q = number of query pairs in $pairs_s$

$\quad\quad$ **for each** pair $< q_1, q_2 >$ in $pairs_s$

$\quad\quad\quad\quad$ $patterns$ = get_shortest_paths(q_1, q_2)

$\quad\quad\quad\quad$ n_p = number of patterns in $patterns$

$$\textbf{for each } \text{pattern } p \text{ in } patterns$$
$$freq\_session_p \mathrel{+{=}} 1/(n_q \cdot n_p)$$

$all\_within\_session\_patterns =$ set of patterns p for which $freq\_session_p$ is defined

for each pattern p in $all\_within\_session\_patterns$

$$support_p = freq\_session_p / \sum_{p' \text{ in } all\_within\_session\_patterns} freq\_session_{p'}$$

Step 3: count patterns between random query pairs

for each pair $< q_1, q_2 >$ in $random\_query\_pairs$

$patterns = \text{get\_shortest\_paths}(q_1, q_2)$

$n_p = $ number of patterns in $patterns$

for each pattern p in $patterns$

$$freq\_random_p \mathrel{+{=}} 1/n_p$$

$all\_random\_session\_patterns =$ set of patterns p for which $freq\_random_p$ is defined

for each pattern p in $all\_random\_session\_patterns$

$$support\_random_p = freq\_random_p / \sum_{p' \text{ in } all\_random\_session\_patterns} freq\_random_{p'}$$
$$confidence_p = support_p / (support_p + support\_random_p)$$

Step 4: iteratively improve the identified patterns

$all\_patterns =$ set of patterns p for which $support_p$ is defined

$new_taboo_patterns_found = $ false

for each pattern p in $all\_patterns$

if $support_p \geq support\_threshold$ **and** $confidence_p < confidence\_threshold$

add p to $taboo\_patterns$

$new_taboo_patterns_found = $ true

if $new_taboo_patterns_found = $ true

repeat steps 2, 3, and 4

else

for each pattern p in $all\_patterns$

add $support_p$ to $modification{-}pattern{-}support$

add $confidence_p$ to $modification{-}pattern{-}confidence$

Function get_shortest_paths

Input: $< q_1, q_2 >$: pair of queries

Output: $abstract\_paths$: list of abstract shortest paths between q_1 and q_2

$entities_{q1} =$ set of entities matching q_1 according to $query{-}entities$

$entities_{q2} =$ set of entities matching q_2 according to $query{-}entities$

$max\_path\_length = 4$

$cur\_path\_length = 1$

$abstract\_paths = \emptyset$

while $abstract\_paths = \emptyset$ **and** $cur\_path\_length \leq max\_path\_length$

$paths =$ all series of cur_path_length triples from $triples$ that connect an entity from $entities_{q1}$ to an entity from $entities_{q2}$. Triples using equivalence relations (`skos:exactMatch` and `own:sameAs` and their subproperties) are not counted in the number of triples.

> **for each** path p in *paths*
>> $p_{abs} = $ extract_predicates(p)
>> **if** there is no pattern p_{taboo} in *taboo_patterns* such that $p_{abs} = p_{taboo}$
>>> add p_{abs} to *abstract_paths*
>
> $cur\_path\_length = cur\_path\_length + 1$

References

(2004). RDF vocabulary description language 1.0: RDF schema. `http://www.w3.org/TR/rdf-schema/`. Last accessed November 5, 2009.

(2009). The DBpedia ontology. `http://wiki.dbpedia.org/Ontology`. Last accessed June 24, 2009.

(2010). OpenCalais Web Service API. `http://www.opencalais.com/calaisAPI`. Last accessed April 28, 2010.

(2010). The Linking Open Data project. `http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`. Last accessed May 12, 2010.

Bar-Ilan, J., Zhu, Z., and Levene, M. (2009). Topic-specific analysis of search queries. In *Proceedings of the 2009 workshop on Web Search Click Data, Barcelona, Spain*, pages 35–42.

Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chowdhury, A., and Frieder, O. (2007). Automatic classification of web queries using very large unlabeled query logs. *ACM Transactions on Information Systems*, 25(2):Article no. 9.

Berners-Lee, T. (2006). Linked data: Design issues. `http://www.w3.org/DesignIssues/LinkedData.html`. Last accessed November 5, 2009.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems, Special Issue on Linked Data*, 4(3):1–22.

Boldi, P., Bonchi, F., Castillo, C., and Vigna, S. (2009). From 'dango' to 'japanese cakes': Query reformulation models and patterns. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy*, pages 183–190.

Bozzon, A., Chirita, P. A., Firan, C. S., and Nejdl, W. (2007). Lexical analysis for modeling web query reformulation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands*, pages 739–740.

Bruza, P. and Dennis, S. (1997). Query reformulation on the internet: Empirical data and the hyperindex search engine. In *Proceedings of the RIAO'97 Conference on Computer-Assisted Searching on the Internet, Montreal, Canada*, pages 488–499.

Costa, R. P. and Seco, N. (2008). Hyponymy extraction and web search behavior analysis based on query reformulation. In *Proceedings of the 11th Ibero-American Conference on AI, Lisbon, Portugal*.

Efthimiadis, E. N. (2000). Interactive query expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database.* MIT press, Cambridge, MA, USA.

Goodrum, A. and Spink, A. (2001). Image searching on the Excite web search engine. *Information Processing and Management*, 37(2):295–311.

He, D., Göker, A., and Harper, D. J. (2002). Combining evidence for automatic web session identification. *Information Processing and Management*, 38(5):727–742.

Hemayati, R., Meng, W., and Yu, C. (2007). Semantic-based grouping of search engine results using WordNet. In *Proceedings of the Eight International Conference on Web-Age Information Management, Hong Kong, China*, pages 678–686. LNCS (Vol. 4505).

Hollink, L., Schreiber, G., and Wielinga, B. (2007). Patterns of semantic relations to improve image content search. *Journal of Web Semantics*, 5(3):195–203.

Hollink, L., Schreiber, G., Wielinga, B., and Worring, M. (2004). Classification of user image descriptions. *International Journal of Human Computer Studies*, 61(5):601–626.

Hollink, V., Tsikrika, T., and De Vries, A. P. (2010a). Semantic vs term-based query modification analysis. In *Proceedings of Tenth Dutch-Belgian Information Retrieval Workshop, Nijmegen, the Netherlands*.

Hollink, V., Tsikrika, T., and De Vries, A. P. (2010b). The semantics of query modification. In *Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010)*.

Huang, J. and Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China*, pages 77–86.

Ingwersen, P. and Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context.* The Information Retrieval Series. The Information Retrieval Series (Vol. 18). Springer.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library and Information Science Research*, 28(3):407–432.

Jansen, B. J. (2008a). The methodology of search log analysis. In Jansen et al. (2008a), chapter 6, pages 1–16.

Jansen, B. J. (2008b). Searching for digital images on the web. *Journal of Documentation*, 64(1):81–101.

Jansen, B. J., Booth, D. L., and Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the Americal Society for Information Science and Technology*, 60(7):1358–1371.

Jansen, B. J., Goodrum, A., and Spink, A. (2000a). Searching for multimedia: analysis of audio, video and image web queries. *World Wide Web*, 3(4):249–254.

Jansen, B. J. and Spink, A. (2006). How are we searching the World Wide Web? a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263.

Jansen, B. J., Spink, A., and Pedersen, J. O. (2004). The effect of specialized multimedia collections on web searching. *Journal of Web Engineering*, 3(3–4):182–199.

Jansen, B. J., Spink, A., and Saracevic, T. (2000b). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227.

Jansen, B. J., Spink, A., and Taksa, I., editors (2008a). *Handbook of Research on Web Log Analysis.* IGI Global.

Jansen, B. J., Taksa, I., and Spink, A. (2008b). Research and methodological foundations of transaction log analysis. In Jansen et al. (2008a), chapter 1, pages 1–16.

Jones, R. and Fain, D. C. (2003). Query word deletion prediction. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, pages 435–436.

Jörgensen, C. and Jörgensen, P. (2005). Image querying by image professionals. *Journal of the Americal Society for Information Science and Technology*, 56(12):1346–1359.

Klyne, G. and Carroll, J. J. (2004). Resource description framework (RDF): Concepts and abstract syntax. `http://www.w3.org/TR/rdf-concepts/`. Last accessed November 5, 2009.

Lau, T. and Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement. In *Proceedings of the Seventh International Conference on User Modeling, Banff, Canada*, pages 119–128.

Lin, W.-H. and Hauptmann, A. G. (2006). Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada*, pages 41–44.

Markkula, M. and Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1:258–285.

Molich, R. and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348.

Nielsen, J. (1994). *Heuristic evaluation*, pages 25–62. John Wiley & Sons.

Özmutlu, H. C. (2009). Markovian analysis for automatic new topic identification in search engine transaction logs. *Applied Stochastic Models in Business and Industry*, 25(6):737–768.

Özmutlu, S. (2006). Automatic new topic identification using multiple linear regression. *Information Processing and Management*, 42(4):934–950.

Özmutlu, S., Özmutlu, H. C., and Spink, A. (2009). Topic analysis and identification of queries. In Jansen et al. (2008a), pages 345–358.

Özmutlu, S., Spink, A., and Özmutlu, H. C. (2003). Multimedia web searching trends: 1997-2001. *Information Processesing and Management*, 39(4):611–621.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Ren, A., Du, X., and Wang, P. (2009). Ontology-based categorization of web search results using YAGO. In *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization, Sanya, China*, pages 800–804.

Rieh, S. Y. and Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management*, 42(3):751–768.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., and Yang, Q. (2006). Query enrichment for web-query classification. *ACM Transactions on Information Systems*, 24(3):320–352.

Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234.

The Getty Foundation (2004a). The art and architecture thesaurus (AAT). `http://www.getty.edu/research/conducting\_research/vocabularies/aat/`.

The Getty Foundation (2004b). The thesaurus of geographical names (TGN). `http://www.getty.edu/research/conducting\_research/vocabularies/tgn/`.

Tjondronegoro, D., Spink, A., and Jansen, B. J. (2009). A study and comparison of multimedia web searching: 1997–2006. *Journal of the American Society for Information Science and Technology*, 60(9):1756–1768.

Van Assem, M., Gangemi, A., and Schreiber, G. (2006). RDF/OWL representation of WordNet, W3C working draft 19 june 2006. `http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/`.

Vossen, P., Maks, I., Segers, R., and Van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech, Morocco*. Used version: Cornetto database version 2.1.

Westman, S. and Oittinen, P. (2006). Image retrieval by end-users and intermediaries in a journalistic work context. In *Proceedings of the First Symposium on Information Interaction in Context, Copenhagen, Denmark*, pages 102–110.

Whittle, M., Eaglestone, B., Ford, N., Gillet, V. J., and Madden, A. (2007). Data mining of search engine logs. *Journal of the American Society for Information Science and Technology*, 58(14):2382–2400.