HIERARCHICAL TAXONOMY EXTRACTION BY MINING TOPICAL QUERY SESSIONS

Miguel Fernández-Fernández
MVConsultoría, Paseo de la Castellana 91, 28046 Madrid, Spain

Daniel Gayo-Avello
University of Oviedo, Despacho 57, Edificio de Ciencias, C/Calvo Sotelo s/n 33007 Oviedo, Spain

Keywords: Web search, Query log, Hyponymy relations, Query reformulation, Automatic taxonomy extraction.

Abstract: Search engine logs store detailed information on Web users interactions. Thus, as more and more people use search engines on a daily basis, important trails of users common knowledge are being recorded in those files. Previous research has shown that it is possible to extract concept taxonomies from full text documents, while other scholars have proposed methods to obtain similar queries from query logs. We propose a mixture of both lines of research, that is, mining query logs not to find related queries nor query hierarchies but actual term taxonomies. In this first approach we have researched the feasibility of finding hyponymy relations between terms or noun-phrases by exploiting specialization search patterns in topical sessions, obtaining encouraging preliminary results.

1 INTRODUCTION

Almost half (49%) of the Internet users in the United States use search engines on a typical day (Fallows, 2008) . This reflects the fact that Web search is becoming a common habit among users. Because of this, the amount of data in query logs is steadily increasing every day, thus, recording a great deal of the common knowledge of the users. As (Paşca, 2007c) pointed out: *"If knowledge is generally prominent or relevant, people will (eventually) ask about it."*

However, searching is not a straightforward process, instead, the users gradually refine both their queries and their goals in a process referred by (Spink et al., 1998) as the successive search phenomenon. During this iterative process the users provide successive queries revealing different search patterns (He et al., 2002). The most relevant ones for this proposal are the so-called Specialization, Generalization, Reformulation, and New. The first pattern, Specialization, refers to the fact that the query q_{i+1} deals with the same topic that q_i but seeks more specialized information (e.g. additional terms have been added to the query). Generalization refers to the opposite, the query q_{i+1} is on the same topic that q_i but seeks more general information (e.g. some terms have been removed from the original query). In the Reformulation search pattern both queries are about the same topic but the user has both added some terms and removed others from the first query and both queries still have some common terms. The last search pattern, New, implies that the queries have not any common term which does not necessarily mean that they are dealing with different topics.

Although such search patterns just rely on lexical information (i.e. the presence or absence of terms) we feel that the number of results satisfying each query can also provide clues about the existence of specialization even when the pattern is just Reformulation (e.g. `dog` and `labrador`, or `ipod` and `electronics`).

In addition to that, it must be noticed that when considering groups of queries we are not interested in all the queries issued by a user during one "sitting" (i.e. a searching episode) but in much shorter fragments where all the queries are topically related. The advantages of using such mini-sessions are two-fold: (1) the data to be considered in order to find semantic relations between terms is much more focused, and (2) such granularity level should dispel most of the privacy issues even if no de-identification was used (Xiong and Agichtein, 2007). In order to obtain such query log segmentation we have employed

Fernández-Fernández M. and Gayo-Avello D. (2009).
HIERARCHICAL TAXONOMY EXTRACTION BY MINING TOPICAL QUERY SESSIONS.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 229-235
DOI: 10.5220/0002331402290235

a technique which has proved to attain similar results to those achieved by a human expert (Gayo-Avello, 2009). Such technique allows us to group topically related queries even when those queries do not share any common term. (see table 1).

Table 1: Five successive records from the AOL query log grouped into a single topical session by our segmentation technique. Please notice that from these queries relations between xenical, xanical (typo), alli, allie (typo) and fat blocker could be obtained.

Session id	Query	Clicked Url
6287652	xanical	http://trustedmeds.com
6287652	allie or xenical	http://stuffonmycat.com
6287652	allie or xenical	http://bangornews.com
6287652	alli fat blocker	http://700club.com
6287652	alli fat blocker	http://wild955.com

2 RELATED WORK

The idea depicted in this paper is somewhat related to previous and on-going works. We will briefly review those most relevant and, then, we will point the main differences between such works and our approach.

First, it must be said that the idea of automatically building term taxonomies is not new and several approaches were proposed to work on full text documents. Works such as (Hearst, 1992) , (Berland and Charniak, 1999), (Caraballo, 1999), (Girju et al., 2003), (Morin and Jacquemin, 2003), among others, are extremely relevant but they cannot be straightforwardly applied to query logs, because most of such techniques require lexico-syntactic patterns and POS tagging which are hardly useful when applied to Web search queries.

With regards to those works relying in query logs (or in folksonomies) there have been two main goals: (1) organizing the queries/tags in hierarchical arrangements (but not actual taxonomies), and (2) automatically obtaining similar queries/tags.

Thus, (Clough et al., 2005) and (Schmitz, 2006) applied subsumption to image tags in order to obtain tag hierarchies. Such hierarchies, however, were not taxonomies because no hyponymy relations were established; instead, the tags were arranged with regards to their specificity (e.g. church ← tower ← bell tower, sanfrancisco ← goldengate). (Heymann and Garcia-Molina, 2006), (Mika, 2007), and (Schwarzkopf et al., 2007) developed rather similar works; they also employed tags collections (although not image tags) and described different techniques to obtain concept hierarchies. Again, such hierarchies were not proper taxonomies.

With regards to the field of query suggestion there exist abundant literature; we will just refer to two recent works that could be confused with our proposal. For instance, (Shen et al., 2008) and (Baeza-Yates and Tiberi, 2007) describe two methods to generate related queries for a given one exploiting the data within the query log; however, neither of such methods produces a proper taxonomy in the form we suggest.

Approaches by other authors could be wrongly considered similar to our approach. For instance, (Chuang and Chien, 2003) describe a method to classify query terms into a predefined category system; thus, it is much closer to query topic classification than to taxonomy bootstrapping. Other works by the same authors such as (Chuang and Chien, 2004) and (Chuang and Chien, 2005), describe methods to obtain term hierarchies but such hierarchies are, in fact, clusters and not taxonomies. There also exist interesting works in the field of information extraction. For instance, (Paşca, 2007a) and (Paşca, 2007c) describe a technique to obtain class attributes from query logs (e.g. finding that population, flag or president are attributes for Country). The same author also provides a method to find named-entities (Paşca, 2007b) which is related to (Sekine and Suzuki, 2007) and (Komachi and Suzuki, 2008). None of these works, however, are related to our approach because they do not generate term taxonomies.

Thus, our proposal, although somewhat related to all the aforementioned research is different in several aspects. Different from classic works –e.g. (Hearst, 1992), (Berland and Charniak, 1999), (Caraballo, 1999), (Girju et al., 2003) and (Morin and Jacquemin, 2003). in that it does not rely in full text documents but in query logs. It also differs from (Clough et al., 2005), (Heymann and Garcia-Molina, 2006), (Schmitz, 2006), (Mika, 2007), (Baeza-Yates and Tiberi, 2007) and (Schwarzkopf et al., 2007) in the underlying goal: while those methods obtain tags or queries hierarchies according to their specificity, we are interested in automatically building actual taxonomies (i.e. hierarchical arrangements according to hyponymy relations). We have also exposed that other works such as (Chuang and Chien, 2003), (Chuang and Chien, 2004), (Chuang and Chien, 2005), (Paşca, 2007c), (Paşca, 2007a), (Paşca, 2007b), (Sekine and Suzuki, 2007) and (Komachi and Suzuki, 2008) are in fact dealing with problems totally unrelated to taxonomy construction.

3 MOTIVATION

By using taxonomies it should be possible to greatly improve search engine results by means of term disambiguation, query suggestion and expansion. However, we feel that current lexical databases (e.g. WordNet (Miller, 1990)) have several issues in order to be really useful for such purposes. First, because WordNet is an English language project, parallel projects for other languages have been developed, such as EuroWordNet (Vossen, 1998), BalkaNet (Greek), Hebrew WordNet, Hindi WordNet and Japanese WordNet among others (Vossen and Fellbaum, 2004). Certainly we could rely on such different wordnets but the task of identifying the language in which queries are written is not trivial given the small number of terms usually employed. Additionally, there exist a huge gap between the lexicon used by Web users and the developers of wordnets. For instance, (Mandala et al., 1999) and (Gabrilovich and Markovitch, 2007) pointed out that most domain-specific relationships between words cannot be found in WordNet, and some kind of words, such as proper names, jargon or slang are just not included. Besides, (Mihalcea, 2003) also explained that due to the fact that professional linguists recognize minimal differences in word senses, common words such as “make” have too many different senses to be useful for IR tasks. Of course, such wordnets could be automatically enriched (Hearst, 1992) but such approach require a great effort (usually carried out by linguists) and, hence, wordnets remain as quite static data sources. On the other hand, most of the aforementioned methods to construct term taxonomies – e.g. (Hearst, 1992), (Berland and Charniak, 1999), (Caraballo, 1999), (Girju et al., 2003), (Morin and Jacquemin, 2003). not only need large text corpora but they are tightly coupled to the grammar rules of the target language. This would make their application to query logs extremely difficult (if not totally unfeasible) given the nature of the queries which are short and, many times, simply ungrammatical. Thus, we feel that taxonomies of terms and noun phrases collecting the common knowledge of search engine users, including typos, jargon and slang are a real need in order to improve the performance of Web search engines. Besides, we think that the only way to obtain such users’ mental model is by mining the query logs collecting the users queries. As a consequence, the following research questions arose: (1) *Is it possible to generate term taxonomies relying only on the queries submitted by search engine users?* And, (2) *Can they be automatically mined in a language-independent way?*

Throughout the following sections we describe our proposal to mine hyponymy relations from query logs, we provide preliminary results from its application to the AOL dataset, in addition to future lines of research.

4 RESEARCH DESIGN

4.1 The AOL dataset

To answer the research questions stated above we employed the AOL query log (Pass et al., 2006). This data set contains more than 30 million records from about 650,000 users sampled from March to May 2006. Each of those records comprises (1) a user identifier, (2) the query string submitted by the user, and (3) the date and time when the query was issued. If the user clicked any result, then the record also includes the host name portion of the clicked URL. As it was previously explained, in addition to that information, we also need the estimated number of results for each query to test the existence of subsumption. Although such information could appear in query logs (e.g. the MSN query log (Microsoft, 2006) does contain it) it is not the case with the AOL data and, hence, we had to enrich the original information by means of the Yahoo! BOSS API[1].

4.2 Data Preparation

We preprocessed the raw log in order to obtain a subset best suited to our purposes. In this phase we sessionized the log and removed those records which were supposed to tamper with the extraction phase. Because of its relevance, that phase will be detailed in a separate section. With regards to the preprocessing phase it comprises the following steps:

1. **Removal of Navigational Queries.** According to (Broder, 2002), there are three kind of queries in relation to their intent: (1)Navigational, when the immediate intent is to reach a particular site; (2) Informational, when the intent is to acquire some information assumed to be present on one or more web pages; and (3)Transactional, when the intent is to perform some web-mediated activity (e.g. to buy a product, or download a file). We think that navigational queries can reduce the accuracy of the hyponym extraction process and, thus, such queries should be removed. Because the intent behind navigational queries is to reach a particular site, most of them are lexically similar

[1] http://developer.yahoo.com/search/boss/

to the URL of the referred site and, thus, a simple heuristic to detect them (Jansen et al., 2008) consists of checking for the appearance of well-known website names (e.g. `google`, `wikipedia`, etc), domain suffixes (e.g. `com`, `net`, ... , `co.uk`, etc), or strings frequently present in URLs such as `www.` or `http://`.

2. **Log Sessionization.** Topical query sessions are sequences of queries issued by a single user and dealing with a unique topic. There are several methods well suited for this task and for these experiments we have applied the technique that, according to the analysis performed in (Gayo-Avello, 2009), achieves the best results. Such a technique relies in both lexical and temporal clues to find a topic shift or a topic continuation between successive queries from the same user.
3. **Removal of Repeated Queries.** Many sessions contain records with repeated queries; there are several reasons for this: maybe the user clicked on more than one search result after submitting the query; s/he could also ask for another page of results; or even the user actually typed the same query more than once during the same session. For our purposes, more than one record with the same value for the query field is simply redundant and, hence, repeated records were removed.
4. **Removal of Low Frequency Queries.** Nonsensical queries or containing odd typos are not rare but they are relatively unfrequent. In order to mitigate the impact of such queries, we removed those records containing queries with an absolute frequency below an arbitrary threshold.
5. **Removal of One-query-sessions.** Because our proposal to detect hyponym relations requires a *specialization* pattern, sessions containing just one query are useless and, therefore, removed.

4.3 Taxonomy Extraction

4.3.1 Identification of Specializations

Our proposal is based on the premise that hyponymy relations can be mined from query logs by taking into account the specialization search pattern in addition to the number of results available for each query.
Specialization occurs when a query q_j looks for information on the same topic than a previous query q_i but it is much more focused. To detect such a pattern researchers have typically relied in lexical similarity [(Miller, 1990), (He et al., 2002)], that is, a query is an specialization when it adds new terms to a previous query. A trivial scenario occurs when the query q_j contains as a substring the previous query q_i (e.g. `fish food` and `tropical fish food`); we will refer to this as *trivial specialization*. A not so trivial case occurs when the subsequent query not only add new terms but remove others from the previous query, such as in the pair `angelica panganiban scandals` and `celebrity scandals`. Arguably, this example is considered an specialization just because the number of terms is larger but, in fact, `angelica panganiban` is a multiword term and, thus, such a pair could also be considered a case of Reformulation. In fact, apart from the appearance of a multiword term, that case would be analogous to the pair comprising `electronic repairs` and `ipod repairs` that clearly fits the Reformulation pattern. It is because of such cases that we propose to use the number of results obtained by each query; that way, if the number of results is significantly different we could consider such reformulations as *reformulations with specialization*, even, if they do not add new terms. Besides, it must be noticed that there exist many queries that do not fit within those interpretations and are, however, specializations (e.g. `labrador` and `dog`). Nevertheless, at this early stage we are only considering trivial specializations and reformulations with specialization. Finally, to increase recall, we have taken every combination of queries within the same topical session instead of just the pairs of successive queries. From such pairs, depending on the type of specialization detected (if any), two different extraction algorithms are applied.

4.3.2 Extraction of Hyponymy Relations

When a reformulation with specialization occurs, we check the number of results obtained by each query in the pair. If such numbers are significantly different we assume that the query with a larger number of results subsumes the other one. Then, the hyponym and hypernym are obtained by removing the terms appearing in both queries (see Figure 1). It must be noticed that this process is not error-free and some noise is, at this moment of the research, unavoidable; for instance, inferring from the queries `pet hamsters` (623K results) and `pet dogs` (4M results) that `hamsters` specializes `dog`; such issues must be addressed in future research.

When dealing with trivial specializations, the previous method is unfeasible because as one of the queries is a substring of the other when subtracting the intersection between both queries from the shorter one it would result in the empty set. We can, of course, directly take the pair as a hyponym relation (e.g. `fish food` ← `tropical fish food`),

Query	Results
book on mother daughter relationships	~17M
trade edition on mother daughter relationships	~261K

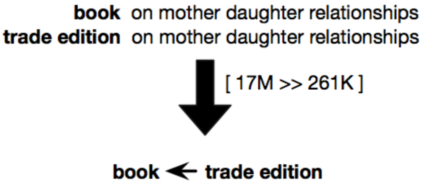
book on mother daughter relationships
trade edition on mother daughter relationships
[17M >> 261K]
book
trade edition

Figure 1: Extraction of a relation from a reformulation with specialization pair. In this case, we would infer that "trade edition" is a hyponym of "book" which is indeed correct (a trade edition is a book intented for general readership)

but other hyponymy relations could also be inferred from the same pattern, such as `fish` ← `tropical fish`, and `food` ← `tropical fish food`. To extract such relations, we employ a method relying on the use of term n-grams. Firstly, we produce for each of the queries every possible n-gram (e.g. for the query `tropical fish food` we would obtain `tropical`, `fish`, `food`, `tropical fish`, `fish food`, and `tropical fish food`). Then, every n-gram of the specialized query is paired with every n-gram of the generalized query provided that (1) both n-grams contain common terms, (2) they are not the same n-gram, and (3) the n-gram from the specialized query is longer than the n-gram from the generalized query. Continuing with the example, the pairs *fish food* and *food*, or *tropical fish* and *fish*, would be obtained, but also *fish food* and *fish* (see Figure 2)

`fish` ← `tropical fish`
`fish` ← `fish food`
`food` ← `fish food`
`fish` ← `tropical fish food`
`food` ← `tropical fish food`
`fish food` ← `tropical fish food`

Figure 2: List of candidate hyponymy relations obtained from the trivial specialization pair (tropical fish food, fish food). As it can be seen, the second and fourth candidates are incorrect.

5 PRELIMINARY RESULTS

We have applied the methods described above to a sessionized and preprocessed version of the AOL query log obtaining encouraging results. At this moment we have not yet developed a way to cuantitatively evaluate the accuracy of the results and, hence, we can only provide a short sample of the relations obtained that we consider illustrative of the pros and cons of our technique:

- `coin` ← `penny` and `military` ← `navy`. These relations are not only correct, but they also appear in Wordnet.
- `lingerie` ← `panties`. Another relationship that should be considered correct although it does not appear in such a straightforward way in Wordnet.
- `celtic` ← `irish`. This relation is highly reasonable and reveals many of the problems of hierarchical taxonomies; that is, are we referring to the Irish language (which indeed belongs to the Celtic family) or to the Irish people?
- `eventing` ← `jumping`. This relation is one the most frequent, appearing in over 300 different sessions. It illustrates very well the way in which specific domains can be exposed because eventing is an equestrian competion comprising several disciplines including jumping; such sense does not appear in Wordnet.
- `motels` ← `howard johnson express` and `wrestling` ← `wwe`. These relations are correct because Howard Johnson Express and WWE are brands dealing with motels and professional wrestling respectively. Because they involve trademarks, they do not appear in Wordnet.
- `underwear` ← `briefs` ← `speedo`. Here we have two hyponymy relations: `underwear` ← `briefs` and `briefs` ← `speedo`. The first relation is correct because briefs are a type of tight underwear and swimwear. The second relation could also be considered correct because, although Speedo is a trademark, it is commonly used as a generic name referring to swimming briefs. Again, these relations reveal common knowledge that is not usually present in lexical databases.
- `mountain` ← `mountian`, `paper` ← `papper`, or `video` ← `vidio` are examples of rather frequent relations mined from the query log. As it can be seen they are not really hyponyms but typos associated with the correct spelling.
- `hanoverian` ← `arabian`, `yellow` ← `white`, `honda` ← `kawasaki`, `justice leage` ← `flash gordon` are examples of some of the issues we have to face up in future research; in these cases we have terms that could be considered co-hyponyms (i.e. terms with a common hypernym) but one of them is much more popular than the other, thus, tampering with our heuristic based on the number of results.

6 CONCLUSIONS AND FUTURE WORK

As it has been exposed above, lexical databases are costly hand made resources that, however, exhibit a lack of common dayly knowledge such as jargon, slang and frequent typos. Nevertheles, such terms, because of their pervasive presence in user Web search queries, are extremely important to improve the performance of search engines. This fact drove us to research the feasibility of automatically extracting term taxonomies from those very same queries. Along this paper we have described an approach with encouraging preliminary results. In fact, it seems that it is not only possible to achieve such results by only using query logs but also that it should be possible to attain that in different languages. Therefore, the stated research questions seem to have a feasible answer

This research also has limitations that should be addressed in the near future. First, a much more precise way to identify specialization patterns is needed. Second, false positives (i.e. incorrectly flagged hyponymy relations) should be filtered out. And third, an evaluation framework should be envisioned in order to quantify the performance of the method. With regards to the first issue we have also pointed out that, at this moment, only lexical clues are employed to detect specialization but we plan to reproduce the work by (Boldi et al., 2009) where they describe a machine learning method to detect much subtle specializations (e.g. `labrador` and `dog`). Regarding the second issue, we have explored a naïve heuristic based on the position where modifiers occur in relation to the hypernym (i.e. they are pre- or post-modifiers). As we pointed before, in the English language such modifiers tend to precede the hypernym (e.g. `tropical fish`, `blue fish`, `recently caught fish`) and, hence, it could be rather simple to remove most of the false positives. This could work in other languages but, certainly, it would not be language independent. However, statistical methods could perhaps be applied to these trivial specializations to discover the most common position of modifiers in order to adapt the application of the heuristic. Finally, with regards to the third issue on the necessity of a evaluation framework, we will probably start relying on Wordnet although we have already pointed out the lack of specialized knowledge and slang in that database. On the other hand, we believe that many pairs would be, in fact, instances and not hyponyms (e.g. `angelina jolie` and `celebrity`) which could be really difficult to evaluate by simply using Wordnet. Hopefully, in future works we will be able to shed light on such issues.

REFERENCES

Baeza-Yates, R. and Tiberi, A. (2007). Extracting semantic relations from query logs. In *KDD '07: Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85, New York, NY, USA. ACM.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., and Vigna, S. (2009). Query suggestions using query-flow graphs. In *WSCD '09: Proc. of the 2009 workshop on Web Search Click Data*, pages 56–63, New York, NY, USA. ACM.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA. Association for Computational Linguistics.

Chuang, S.-L. and Chien, L.-F. (2003). Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decis. Support Syst.*, 35(1):113–127.

Chuang, S.-L. and Chien, L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. In *CIKM '04: Proc. of the thirteenth ACM international conference on Information and knowledge management*, pages 127–136, New York, NY, USA. ACM.

Chuang, S.-L. and Chien, L.-F. (2005). Taxonomy generation for text segments: A practical web-based approach. *ACM Trans. Inf. Syst.*, 23(4):363–396.

Clough, P., Joho, H., and Sanderson, M. (2005). Automatically organising images using concept hierarchies,. In *Proc. of the SIGIR Workshop on Multimedia Information Retrieval*.

Fallows, D. (2008). Almost half of all internet users now use search engines on a typical day. Technical report, Pew Internet & American Life Project. Accessed 6 February 2009. Available at: http://www.pewinternet.org/pdfs//PIP_Search_Aug08.pdf.

Gabrilovich, E. and Markovitch, S. (2007). Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *J. Mach. Learn. Res.*, 8:2297–2345.

Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Inf. Sci.*, 179(12):1822–1843.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

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

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University.

Jansen, B. J., Booth, D. L., and Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266.

Komachi, M. and Suzuki, H. (2008). Minimally supervised learning of semantic knowledge from query logs. In *Proc. of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 358–365.

Mandala, R., Tokunaga, T., and Tanaka, H. (1999). Complementing wordnet with roget's and corpus-based thesauri for information retrieval. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 94–101, Morristown, NJ, USA. Association for Computational Linguistics.

Microsoft (2006). *Microsoft Research Microsoft Live Labs: Accelerating Search in Academic Research 2006*. Available at: http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx. (Accessed 24 November 2008).

Mihalcea, R. (2003). Turning wordnet into an information retrieval resource: Systematic polysemy and conversion to hierarchical codes. *International Journal of Pattern Recognition and Articial Intelligence (IJPRAI)*, pages 689–704.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, pages 235–312.

Morin, E. and Jacquemin, C. (2003). Automatic acquisition and expansion of hypernym links. *Computer and the humanities*, 38:363–396.

Paşca, M. (2007a). Organizing and searching the world wide web of facts – step two: harnessing the wisdom of the crowds. In *WWW '07: Proc. of the 16th international conference on World Wide Web*, pages 101–110, New York, NY, USA. ACM.

Paşca, M. (2007b). Weakly-supervised discovery of named entities using web search queries. In *CIKM '07: Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, pages 683–690, New York, NY, USA. ACM.

Paşca, V. D. (2007c). What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2832–2837.

Pass, G., Chowdhury, A., and Torgeson, C. (2006). A picture of search. In *InfoScale '06: Proc. of the 1st international conference on Scalable information systems*, page 1, New York, NY, USA. ACM.

Schmitz, P. (2006). Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW '06)*.

Schwarzkopf, E., Heckmann, D., Dengler, D., and Kroner, A. (2007). Mining the structure of tag spaces for user modeling. In *Workshop on Data Mining for User Modeling*.

Sekine, S. and Suzuki, H. (2007). Acquiring ontological knowledge from query logs. In *WWW '07: Proc. of the 16th international conference on World Wide Web*, pages 1223–1224, New York, NY, USA. ACM.

Shen, D., Qin, M., Chen, W., Yang, Q., and Chen, Z. (2008). Mining web query hierarchies from clickthrough data. In *AAAI07: Proc. of the Twenty-Second Conference on Artificial Intelligence*.

Spink, A., Wilson, T., Ellis, D., and Ford, N. (1998). Modeling users' successive searches in digital environments. *D-Lib Magazine*. Accesed 6 February 2009. Available at: http://www.dlib.org/dlib/april98/04spink.html.

Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Vossen, P. and Fellbaum, C. (2004). Wordnets in the world. Technical report, Global WordNet Association [http://www.globalwordnet.org/]. Accessed 06-02-09.

Xiong, L. and Agichtein, E. (2007). Towards privacy-preserving query log publishing'. In *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW2007)*.