UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

AUTOMATIC GENERATION OF A LARGE SCALE SEMANTIC SEARCH EVALUATION DATA-SET

Uladzimir Kharkevich

May 2009

Technical Report # DISI-09-031

Also: in the proceedings of the 2nd International Conference on the Semantic Web and Digital Libraries (ICSD), 2009

Automatic Generation of a Large Scale Semantic Search Evaluation Data-Set

Uladzimir Kharkevich

Department of Information Engineering and Computer Science
University of Trento, Italy
kharkevi@disi.unitn.it

Abstract. To compare the performance of information retrieval techniques in various settings, the data-sets which model these settings need to be generated. Although there are already available collections, such as those used in TREC conference series, which are used for evaluation of various retrieval tasks, there is a lack of collections which are specially developed for evaluation of the effectiveness of semantically enhanced text retrieval techniques. In this paper, we propose an approach for the automatic generation of such data-sets, by using search engines query logs and data from human-edited web directories. The evaluation is performed by comparing the performance of *Lucene*, a popular syntactic search engine, and *Concept Search*, a search engine which extends Lucene's syntactic search with semantics.

1 Introduction

The conference series like TREC provide different manually built data-sets for evaluation of search systems performance on various information retrieval tasks. Let us recall some examples of TREC tasks, which are used for evaluation of free-text retrieval. The *Ad Hoc* task examines the performance of systems where the set of documents is fixed and the query set is not known before the experiment. The *Spoken Document Retrieval* task is an *ad hoc* which examines the performance of systems on texts produced by speech recognition systems. The *Terabyte* track, is the *ad hoc* task examining the performance of search systems on large data-sets. All these tasks focus on different information retrieval problems, such as, dealing with corrupted texts and scaling to the big number of documents. The *'Semantic'* task, i.e., the one which would concentrate on the evaluation of the performance of semantically enhanced search systems is missing. The need for a publicly available test corpus for the evaluation of semantic search algorithms is recognized in the semantic search community [12].

In this paper, we first, propose an approach for automatic generation of information retrieval data-sets based on search engines query logs and data from human-edited web directories, and then describe how the data-set for the *'Semantic'* task can be created. This paper is organized as follows. In Section 2, we discuss an approach for automatic generation of information retrieval data-sets. In Section 3, we discuss how the size of the document descriptions can affect the

performance of search techniques. In Section 4, we discuss how *Semantic Heterogeneity*, i.e., one of the main problem which needs to be addressed by semantic search, can be modelled in the data-set. We also discuss how the performance of search techniques is affected by the semantic heterogeneity problem. In Section 5, we discuss the related work in automatic data-set generation. Section 6 concludes the paper.

2 Data-Set Generation

The goal of an *IR* system is to map a natural language query q (in a query set Q), which specifies a certain user information needs, to a set of documents d in the document collection D which meet these needs, and to order these documents according to their relevance (R).

$$IR : Q \xrightarrow{R} D \tag{1}$$

In order to evaluate the efficiency of an *IR* system, we need a data-set which consists at least of the following three components:

Documents (D): Traditionally, documents are represented as Natural Language (NL) texts which vary in size, use different vocabularies, and are about different subject matters. Since most of the real *IR* systems need to deal with large document collections, the set of documents in the data-set should be also big enough. Otherwise, the results obtained on the data-set can not be considered as a good approximation of the real performance of the evaluated system.

Queries (Q): Queries are short statements of user information needs. In fact, the average size of queries which are submitted to the current search engines is less than three words. Such short queries can be ambiguous. In order to be able to evaluate the quality of the results returned by an *IR* system, the data-set should provide an unambiguous description for these queries. For instance, each query in the ad-hoc TREC document collection[1] is assigned a description, i.e., one sentence which describes a topic area, and a narrative, i.e., a concise description of what makes a document relevant to the query.

Relevance judgments (R): A relevance judgment is a query-document pair where the relevance of the document to the query is specified. For instance, in TREC, the binary relevance judgment is used, i.e., either a document is relevant to the query or it is not. The following rule is used by TREC assessors to evaluate the relevancy of a document to the query. If any information contained in a document can be used to write a report about subject of the query, then the document should be marked as relevant. In ideal case, a set of relevancy judgments should be complete and correct. In reality, the size of document collections make it infeasible to produce the complete set of relevance judgments, and, therefore, some approximation of the relevancy

[1] `http://trec.nist.gov/data/test_coll.html`

judgments set is used instead. For example, the pooling methodology is used in TREC [15] to provide such approximation.

In this paper, we propose an approach for automatic generation of data-sets by using search engines query logs and data from human-edited web directories. We use the AOL query log [14], which consists of over 20,000,000 queries made by over 500,000 AOL users during a three-month period. For every query, the query log also contains the time and the sites that were visited by the user. As a web directory we use the *Open Directory Project*[2] (*ODP*), also known as *DMoz*. *DMoz* is a multilingual open content directory of World Wide Web links that is constructed and maintained by a community of more than 80,000 volunteer editors. The *DMoz* directory contains over 590,000 multilingual categories organized into a hierarchy and over 4,500,000 web-sites classified to these categories. The meaning of each category is defined by its positions in the hierarchy. For instance, category *Languages*, which can be reached by a path *Top/Computers/Programming/Languages/* represents a set of web-sites about programming languages and directly related topics. Moreover, all the sub-categories of this category also need to be related to *programming languages*. For instance, category *Java* with the path *Top/Computers/Programming/Languages/Java/* is about programming language Java. Each web-site, in *Dmoz*, is represented by an URL, a title, and a short description of its content. Web-sites are classified to categories according to the *get-specific rule*, i.e., the category which describe the content of the web-site in the most specific way should be chosen. In the following, we discuss how *AOL* query log and *DMoz* web directory can be used for automatic data-set generation.

The documents, in the data-set, are created by using web-sites classified in *DMoz*. First, we collect all the URLs of web-sites classified in DMoz. Note, that we excluded from consideration all the web sites classified in *Adult*, *World*, *Regional* and *Kids_and_Teens* sub-trees. *Adult* sub-tree is excluded because it can contain web-sites with inappropriate adult content, *World* sub-tree is excluded because it contain web-sites with non-English content, and both *Kids_and_Teens* and *Regional* sub-trees are excluded because they have guidelines which are different from those for the rest of the directory. Second, for every URL, we fetch a single web-page pointed by the URL. Third, for every web-page we extract out-links, i.e., URLs which appear in the web-page together with their anchor, and, if there is an out-link with the phrase *about us* (or *about me*), we fetch the web-page corresponding to this URL. All the markup is eliminated from first and *about us* web pages. The fetching of web-site contents and elimination of the markup is implemented by using Nutch[3].

In this paper, as a document set we use only those web-sites which have 'about us' web-pages. We use *AboutUs* as a name for the data-set. Every document in the *AboutUs* data-set consists of three textual fields, which describe what the corresponding web-site is about:

[2] `http://www.dmoz.org/`
[3] `http://lucene.apache.org/nutch/`

Description In DMoz, for each web-site, there is a short description written by DMoz editors which describes what the web-site is about from their point of view. "*The description gives specific information about the content and/or subject matter of the site. It should be informative and concise, usually no longer than one or two lines. The basic formula for a good description is: Description = Subject + Content. ... End users should be able to determine relevancy without having to visit a site.*"[4]

First page First page is the first (and probably the last) think that user see when she visits the web-site. So, the first page should usually give a good idea about web-site content. We see the first page as a description of what the web site is about from the point of view of a web-site visitors.

'About us' Web-site's about page describes what the web site is about from the point of view of web-site authors.

Note, that other web-pages, which can be reached from the first page, can also be used to describe the topic and the content of the web-site. The problem is that it is hard to distinguish between these pages and the ones which are completely unrelated.

In order to generate a query set, we, first, collect all the unique queries from AOL query log. One word queries, queries which contain punctuation, special symbols, or boolean operators (e.g., '+', and '?'), and queries which contain the words shorter than 3 letters are filtered out. Second, for every query, we search for a set **C** of DMoz categories with titles consisting only of the query words (we used exact matching of lowercased words). For example, for query *africa scuba diving* we find categories *Africa* and *Scuba_Diving*. Third, for every category in **C**, we check if its path to the root contains a combination of categories (which are also in **C**), which all together contain all and only query words. For example, the path to the root *Top/Recreation/Outdoors/Scuba_Diving/Regional/Africa* has two categories *Scuba_Diving* and *Africa* with all and only words from the given query. In order to have queries with only one possible interpretation, we filtered out all the queries which matched more than on paths to the root. In Table 1, we show some examples of query-category pairs which we obtained as a result of the described above process. Notice, that many categories in DMoz are assigned descriptions. These descriptions, similarly to the query descriptions in TREC collections, can be used to describe the meaning of the query in the corresponding query-category pair.

In order to generate a set of relevance judgments, we used a mapping from queries to categories obtained as described above and also a mapping from categories to the documents classified to these categories by DMoz editors. For every category, we collect all the documents classified to this category plus all the documents classified to more specific categories. All the documents collected for a category are considered to be relevant to the query in the corresponding query-category pair. Here, the intuition is that, since documents in *DMoz* are sub-categorized and organized by topics[5], all the documents classified in the sub-

[4] `http://www.dmoz.org/guidelines/describing.html#descriptions`

[5] `http://www.dmoz.org/guidelines/subcategories.html`

Table 1. Query-category pairs

AOL Query	DMoz Category
africa scuba diving	Top/Recreation/Outdoors/Scuba_Diving/Regional/Africa
analytical chemistry	Top/Science/Chemistry/Analytical
breast cancer organizations	Top/Health/Conditions_and_Diseases/Cancer/Breast/Organizations
business awards	Top/Business/Consumer_Goods_and_Services/Awards
home based business opportunities	Top/Business/Opportunities/Home_Based
homebrewing beer	Top/Recreation/Food/Drink/Beer/Homebrewing
knowledge management	Top/Reference/Knowledge_Management
laser toner	Top/Computers/Hardware/Peripherals/Printers/Supplies/Laser_Toner
lions clubs international	Top/Society/Organizations/Service_Clubs/Lions_Clubs_International
luxury jewelry	Top/Shopping/Jewelry/Watches/Luxury
nuclear magnetic resonance	Top/Science/Chemistry/Nuclear_Magnetic_Resonance
photography education	Top/Arts/Photography/Education
rehabilitation medicine	Top/Health/Medicine/Medical_Specialties/Rehabilitation_Medicine
rugby football union	Top/Sports/Football/Rugby_Union
shih tzu breeders	Top/Recreation/Pets/Dogs/Breeds/Toy_Group/Shih_Tzu/Breeders
small business accounting software	Top/Computers/Software/Accounting/Small_Business
solar energy business	Top/Business/Energy/Renewable/Solar
travel around the world	Top/Recreation/Travel/Travelogues/Around_the_World
united states adoption	Top/Home/Family/Adoption/Wish_to_Adopt/Regional/United_States
yellow pages directories	Top/Reference/Directories/Address_and_Phone_Numbers/Yellow_Pages

tree should be relevant to all the categories on the path to the root, including those categories which are matched by the query words. Trivial query-document matches, i.e., the ones where documents include query as an exact phrase were excluded from the data-set together with corresponding documents. For example, for a query "west highland white terrier", document "The West Highland White Terrier is a small terrier" is considered trivial, because any syntactic or semantic technique can trivially find this document. Moreover, we pruned all the queries which have less than 10 relevant results. The statistics of the resulting *AboutUs* data-set is summarized in Table 2 [6]. Notice that the generated set of

Table 2. *AboutUs* data-set statistics

Statistics category	Value
Documents	100,807
Queries	330
Relevance judgments	8,704
Query length (words), avg.	2.4
Description length (words), avg.	16.0
First page length (words), avg.	485.4
'About Us' page length (words), avg.	473.3

relevance judgments is correct and complete, in the case, when (i) editors do not do mistakes and do not miss relevant documents, and (ii) the document description is rich enough to judge about the relevance of this document to the query.

[6] White space is used as an indication of a separation between words

According to *DMoz* guidelines: "*An effective editor will search and/or browse through the ODP in areas inside and outside his or her top level category to find areas of potential duplication*"[5]. Assuming that most of editors are "effective editors", the set of relevance judgments (obtained by the described above approach) should be a good approximation of the ideal (i.e., complete and correct) set of relevance judgments. The impact of the richness of the document descriptions on the performance of search techniques is studied in the following section.

3 Document Size

In this section, we study how a size of the web-site description, which can be used as a rough indicator of the amount of available information about the web-site, and the level of details in the descriptions, can affect the performance of search techniques. For our experiments, we used two *IR* systems.

The first system is build on top of Lucene[7], an open source IR toolkit used in many search applications[8]. The system is an instantiation of syntactic search, i.e., a syntactic matching of words is used for matching of document and query terms. Standard tokenization and English Snowball stemmer were used for document and query preprocessing. The AND operator was used as a default boolean operator in a query. The second system is *Concept Search (C-Search)* [4]. *C-Search* is an *IR* approach which is based on retrieval models and data structures of syntactic search, but complex concepts expressed in a propositional Description Logic (DL) language [5, 9] (i.e., a DL language without roles) are used instead of words and syntactic matching of words is extended to semantic matching [8] of complex concepts, where semantic matching is implemented by using positional inverted index. It is not always possible to find atomic concepts which correspond to given words [6], therefore, indexing and retrieval in C-Search are performed by using both syntactic and semantic information, e.g., a word itself is used as a concept, when its denoted concept is not known. *C-Search* can be seen as a semantics enabled version of Lucene.

Three data-sets were generated based on the *AboutUs* data-set (see Section 2). These data-sets represent differen levels of details in document description. The first data-set (**descr**) consists only of short descriptions, created by DMoz editors, which briefly describe the web-site. In the second data-set (**descr+fp**), every document is composed from the description and the text from the first page of the web-site. The third data-set (**descr+fp+ap**) consists of the documents, which are composed from description, first and 'about us' web-pages. Actually, **descr+fp+ap** represents the complete *AboutUs* data-set.

We evaluated the performance of Lucene and *C-Search* on all three data-sets. In the evaluation, we used the standard IR measures: (i) the mean average precision (MAP), and (ii) precision at K (P@K), where K was set to 5 and 10. The average precision for a query is the mean of the precision obtained after each relevant document is retrieved (using 0 as the precision for not retrieved

[7] `http://lucene.apache.org/java/docs/index.html`
[8] `http://wiki.apache.org/lucene-java/PoweredBy`

documents which are relevant). MAP is the mean of the average precisions for all the queries in the test collection. P@K is the percentage of relevant documents among the top K ranked documents. MAP is used to evaluate the overall accuracy of IR system, while P@K is used to evaluate the utility of IR system for users who only see the top K results. The evaluation results are reported in Figure 1. Also, in Figure 1, we provide recall-precision graphs, i.e., we plot precision as a function of recall.

	descr			**descr+fp**			**descr+fp+ap**		
	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10
Lucene	0.0200	0.0879	0.0558	0.1008	0.2255	0.1945	0.1349	0.2473	0.2236
C-Search(Lucene)	0.0359	0.1230	0.0924	0.1411	0.2345	0.2182	0.1798	0.2685	0.2524
Improvement	+79.5%	+39.9%	+65.6%	+40.0%	+4.0%	+12.2%	+33.3%	+8.6%	+12.9%

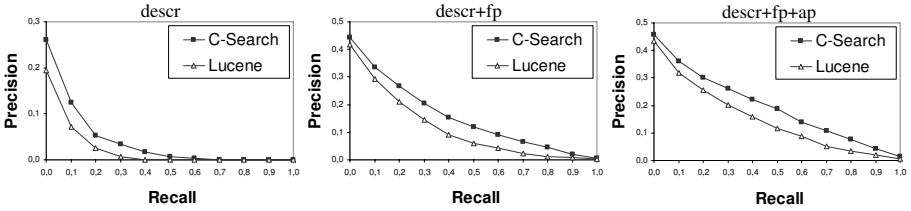

Fig. 1. Evaluation results: Document Size

The experiments show, that, the bigger is the document description, the easier is the search task for both Lucene and *C-Search*. After manual inspection of the results, we concluded that the main reason for this is the increase in the quality of the data-set. If a document description is only a short summary of a web-site (as it is the case in the **descr** data-set), it may not be relevant to a query created for a category in which the web-site is classified. For instance, let us consider the following document description: *Links to auto reviews and articles*. The description is created for the web-site classified to category *New*[9] and, therefore, this document description can be associated with the query *purchasing new automobiles*, but, as we can see, this description contains no information relevant to purchasing of something new. If, in addition to the description, we consider also the first page (as in the **descr+fp** data-set), and 'About Us' page (as in the **descr+fp+ap** data-set) then the web-site description become more complete and the search techniques improve the results. Note, however, that data collected from web-sites can be very noisy, because usually there are many advertisements on web-sites and/or because web-site administrators use different search engine optimization (SEO) techniques, such as adding popular keywords to their web-pages in order to improve the find-ability of their web-sites. In general, it can cause decrease in precision. As we can observe from Figure 1, the completeness of the document descriptions and the noisiness of web-pages are

[9] `http://www.dmoz.org/Home/Consumer_Information/Automobiles/Purchasing/New/`

not playing decisive role if we want to conduct comparative evaluation of different search techniques. For example, C-Search performs better than Lucene on all three data-sets.

4 Semantic Heterogeneity

In the context of IR, semantic heterogeneity refers to a phenomenon, when a person submitting a search query and authors of documents have no agreement about how to represent the same or related objects. For instance, it can lead to the situation, when words which are used to describe the object in a query are different from those words which are used to describe the same object in the document description. In this section, we study how the semantic heterogeneity problem can affect the performance of search techniques.

We create three data-sets: **descr+fp+ap_25**, **descr+fp+ap_10**, and **descr+fp+ap_0**, which are based on *AboutUs* data-set (**descr+fp+ap**). The number X, which appears at the end of the data-set name **descr+fp+ap_X**, represents the percentage of relevant documents which can have all the words from the corresponding query. The data-sets were created by excluding all the documents and corresponding relevance judgments which were above the specified limit. Notice, that the bigger is X, the higher is the level of semantic heterogeneity, where the **descr+fp+ap_0** data-set represents the extreme case when syntactic search is not possible. The performance of Lucene and *C-Search* was evaluated on these data-sets. The evaluation results are reported in Figure 2.

	descr+fp+ap_25			**descr+fp+ap_10**			**descr+fp+ap_0**		
	MAP	P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10
Lucene	0.1098	0.2188	0.1724	0.0466	0.1497	0.1064	0.0000	0.0000	0.0000
C-Search(Lucene)	0.1543	0.2400	0.2079	0.1178	0.1824	0.1496	0.0604	0.0804	0.0707
Improvement	+40.5%	+9.7%	+20.6%	+152.8%	+21.8%	+40.7%	-	-	-

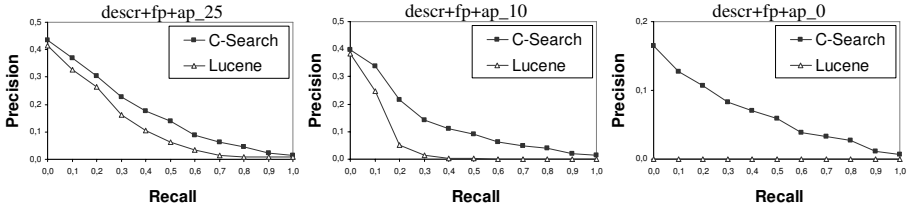

Fig. 2. Evaluation results: Semantic Heterogeneity

From Figure 2, we observe that improvements, achieved by *C-Search*, starts being significant when the heterogeneity is high (i.e., when the number X is small). In order to compare the level of semantic heterogeneity in the generated data-sets with those in standard *IR* data-sets, we took three TREC data-sets: TREC6 (topics 301-350), TREC7 (topics 351-400), and TREC8 (topics 401-450).

The average number (and the average percentage) of relevant documents which have all the query words is computed for these data-sets (see Table 3).

Table 3. Semantic heterogeneity in TREC ad-hoc data-sets

Data-set	Average number of relevant documents which contain all the query words	Average percentage of relevant documents which contain all the query words
TREC6	23.9	27.32 %
TREC7	24.2	29.67 %
TREC8	34.7	40.98 %

As we can see from Table 3, in TREC data-sets, more than 20 relevant documents in average can be retrieved by syntactic matching of document and query words. These documents in average amount to more than 25% of all the relevant documents. The level of semantic heterogeneity problem in TREC data-sets is rather low to show the advantages of semantic techniques (especially when retrieval of top-k results is considered).

The set of relevance judgments, in the **descr+fp+ap_0** data-set, consists only of query-document pairs, where documents and queries are related according to their meanings (and not syntax). Therefore, the **descr+fp+ap_0** data-set can be specifically used for evaluating the effectiveness of semantically enhanced text retrieval techniques.

5 Related work

The need for (semi-)automatic approaches to the data-set generation was recognized in different areas of Information Retrieval and Semantic Web. These areas include text categorization [1, 3, 10], ontology matching [2, 7], and web search [11, 13]. Human-edited web directories are often used as a valuable source of the real data for creating the data-sets. For instance, in [7], an ontology matching evaluation data-set *TaxME2*, composed of thousands of atomic matching tasks, is build semi-automatically out of the Google, Yahoo and Looksmart web directories. The methodology for automatically acquiring labelled data-sets for text categorization is described in [3]. The knowledge encoded into the structure of the DMoz web directory is used in order to generate numerous data-sets with desired properties. Various metrics which can be used in order to estimate the difficulty of the created data-sets are discussed in [3]. The DMoz web-directory and the AOL query log are also used in [13] for automatic evaluation of effectiveness of web search engines. Differently from our approach, query-category pairs are created only for categories whose labels exactly matched the query. Instead, in our approach the path to the root is also analyzed. Moreover, because of the specificity of the search task, construction of the documents is not discussed in [13]. More importantly, in [13], the dynamics of web search is studied, whereas, we are concentrating on the evaluation of semantic search techniques.

6 Conclusion

In this paper, we presented an approach for automatic generation of the data-sets for evaluation of semantics enabled free-text search. The generation of the data-sets is done using search engine query logs and data from human-edited web directories. Future work includes:

- The quality of generated data-sets needs to be evaluated. As a result, some filtering mechanisms might need to be introduced. For example, in [3], it is proposed to employ filtering mechanisms before, during, and after downloading the data from web-pages.
- In our approach, queries are created from categories in the web directory. The categories, apart from descriptions, can have additional context encoded into their positions in the category hierarchy. We need to study how this contextual information can be used by semantic techniques in order to improve their performance. For example, the context can be used in order to disambiguate the meaning of words in the query [16].
- In DMoz, to improve navigability, apart from the main hierarchical structure, additional links are created between related categories. Examples of such links are *@link* and *related* links. *@links*, for example, "*are used to link from one category to another that could theoretically be a subcategory of the first*"[10] These links can be seen as some sort of semantic links. We need to study if/how they can be used for the data-set generation.
- The binary relevance judgment does not represent the importance of the document to the query with respect to other relevant documents. Therefore, all the relevant documents are considered to be equivalent. We need to study how the link structure of DMoz can be used to produce a more accurate judjments. For example, some kind of hierarchical distance can be used [11].
- The size of the data-set can be increased. Since for the evaluation we need only the positive relevance judgments, we can increase the number of documents in the data-set by adding all the web-sits in the English part of DMoz. Queries, in general, can be created from any category and not only from queries in the query log.

Acknowledgments. This work has been partially supported by the European Commission - LIVING KNOWLEDGE project, Grant agreement no.: 231126. Thanks to Fausto Giunchiglia for useful discussions and feedback on this paper.

References

1. P. Avesani, C. Girardi, N. Polettini, and D. Sona. TaxE: a testbed for hierarchical document classifiers. Technical Report T04-04-02, ITC-IRST, 2004.
2. Paolo Avesani, Fausto Giunchiglia, and Mikalai Yatskevich. A large scale taxonomy mapping evaluation. In *4th International Semantic Web Conference (ISWC)*. Springer, 2005.

[10] `http://www.dmoz.org/erz/categories/relcats.html`

3. Dmitry Davidov, Evgeniy Gabrilovich, and Shaul Markovitch. Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–257. ACM, 2004.
4. Fausto Giunchiglia, Uladzimir Kharkevich, and Ilya Zaihrayeu. Concept search. In *Proc. of ESWC'09*, Lecture Notes in Computer Science. Springer, 2009.
5. Fausto Giunchiglia, Maurizio Marchese, and Ilya Zaihrayeu. Encoding classifications into lightweight ontologies. In *Journal on Data Semantics (JoDS) 8*, 2006.
6. Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Discovering missing background knowledge in ontology matching. In *Proc. of ECAI*, 2006.
7. Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani, and Pavel Shvaiko. A large scale dataset for the evaluation of ontology matching systems. *The Knowledge Engineering Review Journal*, 2008.
8. Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic matching: Algorithms and implementation. *Journal on Data Semantics (JoDS)*, 9:1–38, 2007.
9. Fausto Giunchiglia and Ilya Zaihrayeu. Lightweight ontologies. In *The Encyclopedia of Database Systems*, 2008.
10. Fausto Giunchiglia, Ilya Zaihrayeu, and Uladzimir Kharkevich. Formalizing the get-specific document classification algorithm. In *ECDL*. Springer, 2007.
11. Taher H. Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. Evaluating strategies for similarity search on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*. ACM, 2002.
12. M. Hildebrand, J. van Ossenbruggen, and L. Hardman. An analysis of search-based user interaction on the semantic web. Technical Report INS-E0706, CWI, 2007.
13. Eric C. Jensen, Steven M. Beitzel, Abdur Chowdhury, and Ophir Frieder. Repeatable evaluation of search services in dynamic environments. *ACM Trans. Inf. Syst.*, 2007.
14. Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *InfoScale'06: Proceedings of the 1st international conference on Scalable information systems*, New York, NY, USA, 2006. ACM.
15. Ellen M. Voorhees. Overview of trec 2006. In *Proceedings of the 15th Text REtrieval Conference (TREC)*, 2006.
16. I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang. From web directories to ontologies: Natural language processing challenges. In *6th International Semantic Web Conference (ISWC 2007)*. Springer, 2007.