Semantic Search on Text and Knowledge Bases

Hannah Bast
University of Freiburg
bast@cs.uni-freiburg.de

Björn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Elmar Haussmann
University of Freiburg
haussmann@cs.uni-freiburg.de
Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2016, Volume 10, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/1500000032
Semantic Search on Text and Knowledge Bases

Hannah Bast
University of Freiburg
bast@cs.uni-freiburg.de

Björn Buchhold
University of Freiburg
buchhold@cs.uni-freiburg.de

Elmar Haussmann
University of Freiburg
haussmann@cs.uni-freiburg.de
# Contents

1 Introduction .................................................. 2
   1.1 Motivation for this Survey ............................... 2
   1.2 Scope of this Survey ................................. 5
   1.3 Overview of this Survey ............................... 8
   1.4 Glossary .............................................. 10

2 Classification by Data Type and Search Paradigm .............. 13
   2.1 Data Types and Common Datasets ..................... 14
   2.2 Search Paradigms .................................. 25
   2.3 Other Aspects ...................................... 30

3 Basic NLP Tasks in Semantic Search .......................... 32
   3.1 Part-of-Speech Tagging and Chunking ................ 33
   3.2 Named-Entity Recognition and Disambiguation ........ 35
   3.3 Sentence Parsing ................................... 40
   3.4 Word Vectors ....................................... 44

4 Approaches and Systems for Semantic Search ................. 49
   4.1 Keyword Search in Text ............................. 50
   4.2 Structured Search in Knowledge Bases ............... 55
   4.3 Structured Data Extraction from Text ............... 61
   4.4 Keyword Search on Knowledge Bases ................. 72
Abstract

This article provides a comprehensive overview of the broad area of semantic search on text and knowledge bases. In a nutshell, semantic search is “search with meaning”. This “meaning” can refer to various parts of the search process: understanding the query (instead of just finding matches of its components in the data), understanding the data (instead of just searching it for such matches), or representing knowledge in a way suitable for meaningful retrieval.

Semantic search is studied in a variety of different communities with a variety of different views of the problem. In this survey, we classify this work according to two dimensions: the type of data (text, knowledge bases, combinations of these) and the kind of search (keyword, structured, natural language). We consider all nine combinations. The focus is on fundamental techniques, concrete systems, and benchmarks. The survey also considers advanced issues: ranking, indexing, ontology matching and merging, and inference. It also provides a succinct overview of natural language processing techniques that are useful for semantic search: POS tagging, named-entity recognition and disambiguation, sentence parsing, and word vectors.

The survey is as self-contained as possible, and should thus also serve as a good tutorial for newcomers to this fascinating and highly topical field.
1

Introduction

1.1 Motivation for this Survey

This is a survey about the broad field of semantic search. Semantics is the study of meaning. In a nutshell, therefore, it could be said that semantic search is search with meaning.

Let us first understand this by looking at the opposite. Only a decade ago, search engines, including the big web search engines, were still mostly lexical. By lexical, we here mean that the search engine looks for literal matches of the query words typed by the user or variants of them, without making an effort to understand what the whole query actually means.

Consider the query university freiburg issued to a web search engine. Clearly, the homepage of the University of Freiburg is a good match for this query. To identify this page as a match, the search engine does not need to understand what the two query words university and freiburg actually mean, nor what they mean together. In fact, the university homepage contains these two words in its title (and, as a

\footnote{The word comes from the ancient greek word sëmantikós, which means important.}
1.1. **Motivation for this Survey**

matter of fact, no other except the frequent word *of*). Further, the page is at the top level of its domain, as can be seen from its URL: [http://www.uni-freiburg.de](http://www.uni-freiburg.de). Even more, the URL consists of parts of the query words. All these criteria are easy to check, and they alone make this page a very good candidate for the top hit of this query. No deeper understanding of what the query actually “meant” or what the homepage is actually “about” were needed.\(^2\)

Modern search engines go more and more in the direction of accepting a broader variety of queries, actually trying to “understand” them, and providing the most appropriate answer in the most appropriate form, instead of just a list of (excerpts from) matching documents.

For example, consider the two queries *computer scientists* and *female computer scientists working on semantic search*. The first query is short and simple, the second query is longer and more complex. Both are good examples of what we would call semantic search. The following discussion is independent of the exact form of these queries. They could be formulated as keyword queries like above. They could be formulated in the form of complete natural language queries. Or they could be formulated in an abstract query language. The point here is what the queries are asking for.

To a human, the intention of both of these queries is quite clear: the user is (most likely) looking for scientists of a certain kind. Probably a list of them would be nice, with some basic information on each (for instance, a picture and a link to their homepage). For the query *computer scientists*, Wikipedia happens to provide a page with a corresponding list and matching query words.\(^3\) Correspondingly, the list is also contained in DBpedia, a database containing the structured knowledge from Wikipedia. But in both cases it is a manually compiled list, limited to relatively few better-known computer scientists. For the second query (*female computer scientists working on semantic search*), there is no single web page or other document with a corresponding

---

\(^2\)In this simple example, we are leaving aside the important issue of *spam*. That is, someone deliberately putting misleading keywords in the title or even in the URL, in order to fool search engines, and thus users, to consider the web page relevant. Note that this query could also be solved using clickthrough data; see Section 1.2.2.

list, let alone one matching the query words. Given the specificity of the query, it is also unlikely that someone will ever manually compile such a list (in whatever format) and maintain it. Note that both lists are constantly changing over time, since new researchers may join any time.

In fact, even individual web pages matching the query are unlikely to contain most of the query words. A computer scientist does not typically put the words computer scientist on his or her homepage. A female computer scientist is unlikely to put the word female on her homepage. The homepage probably has a section on that particular scientist’s research interests, but this section does not necessarily contain the word working (maybe it contains a similar word, or maybe no such word at all, but just a list of topics). The topic semantic search will probably be stated on a matching web page, though possibly in a different formulation, for example, intelligent search or knowledge retrieval.

Both queries are thus good examples, where search needs to go beyond mere lexical matching of query words in order to provide a satisfactory result to the user. Also, both queries (in particular, the second one) require that information from several different sources is brought together to answer the query satisfactorily. Those information sources might be of different kinds: (unstructured) text as well as (structured) knowledge bases.

There is no exact definition of what semantic search is. In fact, semantic search means a lot of different things to different people. And researchers from many different communities are working on a large variety of problems related to semantic search, often without being aware of related work in other communities. This is the main motivation behind this survey.

When writing the survey, we had two audiences in mind: (i) newcomers to the field, and (ii) researchers already working on semantic search. Both audiences should get a comprehensive overview of which approaches are currently pursued in which communities, and what the current state of the art is. Both audiences should get pointers for further reading wherever the scope of this survey (defined in Section 1.2
right next) ends. But we also provide explanations of the underlying concepts and technologies that are necessary to understand the various approaches. Thus, this survey should also make a good tutorial for a researcher previously unfamiliar with semantic search.

1.2 Scope of this Survey

1.2.1 Kinds of Data

This survey focuses on semantic search on text (in natural language) or knowledge bases (consisting of structured records). The two may also be combined. For example, a natural language text may be enriched with semantic markup that identifies mentions of entities from a knowledge base. Or several knowledge bases with different schemata may be combined, like in the Semantic Web. The types of data considered in this survey are explained in detail in Section 2.1 on Data Types and Common Datasets.

This survey does not cover search on images, audio, video, and other objects that have an inherently non-textual representation. This is not to say that semantic search is not relevant for this kind of data; quite the opposite is true. For example, consider a user looking for a picture of a particular person. Almost surely, the user is not interested in the precise arrangements of pixels that are used to represent the picture. She might not even be interested in the particular angle, selection, or lighting conditions of the picture, but only in the object shown. This is very much “semantic search”, but on a different kind of data. There is some overlap with search in textual data, including attempts to map non-textual to textual features and the use of text that accompanies the non-textual object (e.g., the caption of an image). But mostly, search in non-textual data is a different world that requires quite different techniques and tools.

A special case of image and audio data are scans of text documents and speech. The underlying data is also textual and can be extracted using optical character recognition (OCR) and automatic speech recognition (ASR) techniques. We do not consider these techniques in this

\footnote{Leaving aside aspects like a particular writing style or emotions when talking.}
survey. However, we acknowledge that “semantic techniques”, as described in this survey, can be helpful in the text recognition process. For example, in both OCR and ASR, a semantic understanding of the possible textual interpretations can help to decide which interpretation is the most appropriate.

1.2.2 Kinds of Search

There are three types of queries prevailing in semantic search: keyword, structured, and natural language. We cover the whole spectrum in this survey; see Section 2.2 on Search Paradigms.

Concerning the kind of results returned, we take a narrower view: we focus on techniques and systems that are extractive in the sense that they return elements or excerpts from the original data. Think of the result screen from a typical web search engine. The results are nicely arranged and partly reformatted, so that we can digest them properly. But it’s all excerpts and elements from the web pages and knowledge bases being searched in the background.

We only barely touch upon the analysis of query logs (queries asked) and clickthrough data (results clicked). Such data can be used to derive information on what users found relevant for a particular query. Modern web search engines leverage such information to a significant extent. This topic is out of scope for this survey, since an explicit “understanding” of the query or the data is not necessary. We refer the user to the seminal paper of Joachims [2002] and the recent survey of Silvestri [2010].

There is also a large body of research that involves the complex synthesis of new information, in particular, text. For example, in automatic summarization, the goal is to summarize a given (long) text document, preserving the main content and a consistent style. In multi-document summarization, this task is extended to multiple documents on a particular topic or question. For example, compile a report on drug trafficking in the United States over the past decade. Apart from collecting the various bits and pieces of text and knowledge required to answer these questions, the main challenge becomes to compile these into a compact and coherent text that is well comprehensible for hu-
1.2. **Scope of this Survey**

mans. Such non-trivial automatic content synthesis is out of scope for this survey.

1.2.3 **Further inclusion criteria**

As just explained, we focus on semantic search on text and knowledge bases that retrieves elements and excerpts from the original data. But even there we cannot possibly cover all existing research in depth.

Our inclusion criteria for this survey are very practically oriented, with a focus on fundamental techniques, datasets, benchmarks, and systems. Systems were selected with a strong preference for those evaluated on one of the prevailing benchmarks or that come with a working software or demo. We provide quantitative information (on the benchmarks and the performance and effectiveness of the various systems) wherever possible.

We omit most of the history and mostly focus on the state of the art. The historical perspective is interesting and worthwhile in its own right, but the survey is already long and worthwhile without this. However, we usually mention the first system of a particular kind. Also, for each of our nine categories (explained right next, in Section 1.3), we describe systems in chronological order and make sure to clarify the improvements of the newer systems over the older ones.

1.2.4 **Further Reading**

The survey provides pointers for further reading at many places. Additionally, we provide here a list of well-known conferences and journals, grouped by research community, which are generally good sources for published research on the topic of this survey and beyond. In particular, the bibliography of this survey contains (many) references from each of these venues. This list is by no means complete, and there are many good papers that are right on topic but published in other venues.

**Information Retrieval:** SIGIR, CIKM, TREC, TAC, FNTIR.

**Web and Semantic Web:** WWW, ISWC, ESWC, AAAI, JWS.

**Computer linguistics:** ACL, EMNLP, HLT-NAACL.

**Databases / Data Mining:** VLDB, KDD, SIGMOD, TKDE.
1.3 Overview of this Survey

Section 1.4 provides a Glossary of terms that are strongly related to semantic search. For each of these, we provide a brief description together with a pointer to the relevant passages in the survey. This is useful for readers who specifically look for material on a particular problem or aspect.

Section 2 on Classification by Data Type and Search Paradigm describes the two main dimensions that we use for categorizing research on semantic search:

*Data type*: text, knowledge bases, and combined data.

*Search paradigm*: keyword, structured, and natural language search.

For each data type, we provide a brief characterization and a list of frequently used datasets. For each search paradigm, we provide a brief characterization and one or two examples.

Section 3 on Basic NLP Tasks in Semantic Search gives an overview of: part-of-speech (POS) tagging, named-entity recognition and disambiguation (NER+NED), parsing the grammatical structure of sentences, and word vectors/embeddings. These are used as basic building blocks by various (though not all) of the approaches described in our main Section 4. We give a brief tutorial on each of these tasks, as well as a succinct summary of the state of the art.

Section 4 on Approaches and Systems for Semantic Search is the core section of this survey. We group the many approaches and systems that exist in the literature by data type (three categories, see above) and search paradigm (three categories, see above). The resulting nine combinations are shown in Figure 1.1. In a sense, this figure is the main signpost for this survey. Note that we use Natural Language Search and Question Answering synonymously in this survey. All nine subsections share the same sub-structure:

*Profile* ... a short characterization of this line of research

*Techniques* ... what are the basic techniques used

*Systems* ... a concise description of milestone systems or software

*Benchmarks* ... existing benchmarks and the best results on them
1.3. Overview of this Survey

Figure 1.1: Our basic classification of research on semantic search by underlying data (rows) and search paradigm (columns). The three data types are explained in Section 2.1, the three search paradigms are explained in Section 2.2. Each of the nine groups is discussed in the indicated subsection of our main Section 4.

Section 5 on Advanced Techniques for Semantic Search deals with: ranking (in semantic entity search), indexing (getting not only good results but getting them fast), ontology matching and merging (dealing with multiple knowledge bases), and inference (information that is not directly contained in the data but can be inferred from it). They provide a deeper understanding of the aspects that are critical for results of high quality and/or with high performance.

Section 6 on The Future of Semantic Search provides a very brief summary of the state of the art in semantic search, as described in the main sections of this survey, and then dares to take a look into the near and the not so near future.

The article closes with a long list of 218 references. Datasets and standards are not listed as part of the References but separately in the Appendices. In the PDF of this article, all citations in the text are clickable (leading to the respective entry in the References), and so are

Full text available at: http://dx.doi.org/10.1561/1500000032
Introduction

most of the titles in the References (leading to the respective article on the Web). In most PDF readers, $\text{Alt+Left}$ brings you back to the place of the citation.

The reader may wonder about possible reading orders and which sections depend upon which. In fact, each of the six sections of this survey is relatively self-contained and readable on its own. This is true even for each of the nine subsections (one for each kind of semantic search, according to our basic classification) of the main Section 4. However, when reading such a subsection individually, it is a good idea to prepend a quick read of those subsections from Section 2 that deal with the respective data type and search paradigm: they are short and easy to read, with instructive examples. Readers looking for specific information may find the glossary, which comes right next, useful.

1.4 Glossary

This glossary provides a list of techniques or aspects that are strongly related to semantic search but non-trivial to find using our basic classification. For each item, we provide a very short description and a pointer to the relevant section(s) of the survey.

Deep learning for NLP: natural language processing using (deep) neural networks; used for the word vectors in Section 3.4; some of the systems in Section 4.8 on Question Answering on Knowledge Bases use deep learning or word vectors; apart from that, deep NLP is still used very little in actual systems for semantic search, but see Section 6 on The Future of Semantic Search.

Distant supervision: technique to derive labeled training data using heuristics in order to learn a (supervised) classifier; the basic principle and significance for semantic search is explained in Section 4.3.2 on Systems for Relationship Extraction from Text.

Entity resolution: identify that two different strings refer to the same entity; this is used in Section 4.3.4 on Knowledge Base Construction and discussed more generally in Section 5.4 on Ontology Matching and Merging.
1.4. Glossary

**Entity search/retrieval:** search on text or combined data that aims at a particular entity or list of entities as opposed to a list of documents; this applies to almost all the systems in Section 4 that work with combined data or natural language queries\(^5\); see also Section 5.1, which is all about ranking techniques for entity search.

**Knowledge base construction:** constructing or enriching a knowledge base from a given text corpus; basic techniques are explained in Section 4.3.1; systems are described in Section 4.3.4.

**Learning to rank for semantic search:** supervised learning of good ranking functions; several applications in the context of semantic search are described in Section 5.1.

**Ontology merging and matching:** reconciling and aligning naming schemes and contents of different knowledge bases; this is the topic of Section 5.3.

**Paraphrasing or synonyms:** identifying whether two words, phrases or sentences are synonymous; systems in Section 4.8 on *Question Answering on Knowledge Bases* make use of this; three datasets that are used by systems described in this survey are: Patty [2013] (paraphrases extracted in an unsupervised fashion), Paralex [2013] (question paraphrases), and CrossWikis [2012] (Wikipedia entity anchors in multiple languages).

**Question answering:** synonymous with natural language search in this survey; see Section 2.2.3 for a definition; see Sections 4.7, 4.8, and 4.9 for research on question answering on each of our three data types.

**Reasoning/Inference:** using reasoning to infer new triples from a given knowledge base; this is the topic of Section 5.4.

**Semantic parsing:** finding the logical structure of a natural language query; this is described in Sections 4.8 on *Question Answering on Knowledge Bases* and used by many of the systems there.

**Semantic web:** a framework for explicit semantic data on the web; this kind of data is described in Section 2.1.3; the systems described

\(^5\)A search on a knowledge base naturally returns a list of entities, too. However, the name *entity search* is usually only used when (also) text is involved and returning lists of entities is not the only option.
in Section 4.5 deal with this kind of data; it is important to note that many papers / systems that claim to be about semantic web data are actually dealing only with a single knowledge base (like DBpedia, see Table 2.2), and are hence described in the sections dealing with search on knowledge bases.

**Information extraction:** extracting structured information from text; this is exactly what Section 4.3 on *Structured Data Extraction from Text* is about.

**XML retrieval:** search in nested semi-structured data (text with tag pairs, which can be arbitrarily nested); the relevance for semantic search is discussed in Section 4.5.3 in the context of the INEX series of benchmarks.


