Semantic Search

Philippe Cudre-Mauroux

Definition

Semantic Search regroups a set of techniques designed to improve traditional document or knowledge base search. Semantic Search aims at better grasping the context and the semantics of the user query and/or of the indexed content by leveraging Natural Language Processing, Semantic Web and Machine Learning techniques to retrieve more relevant results from a search engine.

Overview

Semantic Search is an umbrella term regrouping various techniques for retrieving more relevant content from a search engine. Traditional search techniques focus on ranking documents based on a set of keywords appearing both in the user's query and in the indexed content. Semantic Search, instead, attempts to better grasp the semantics (i.e., meaning) and the context of the user query and/or of the indexed content in order to retrieve more meaningful results.

Semantic Search techniques can be broadly categorized into two main groups depending on the target content:

- techniques improving the relevance of classical search engines where the query consists of natural language text (e.g., a list of keywords) and results are a ranked list of documents (e.g., webpages);
- techniques retrieving semi-structured data (e.g., entities or RDF triples) from a knowledge base (e.g., a knowledge graph or an ontology) given a user query formulated either as natural language text or using a declarative query language like SPARQL.

Those two groups are described in more detail in the following section. For each group, a wide variety of techniques have been proposed, ranging from Natural Language Processing (to better grasp the contents of the query and data) to Semantic Web (to guide the search process leveraging declarative artifacts like ontologies) and Machine Learning (typically to learn models from large quantities of data).

Main Approaches

We give below an overview of the various techniques that have been proposed in the context of Semantic Search for improving document search as well as knolwedge base search. A number of surveys delve into more detail in this context: Mangold (2007) focuses on natural language queries on RDF knowledge bases or ontologies. Madhu et al (2011) and Mäkelä (2005) are two brief surveys covering both topics. Bast et al (2016) is an extensive survey covering Semantic Search in its broadest sense.

Document Search

Classical search engines take as input a user query formulated as a list of keywords and return as output a ranked list of documents relevant to those keywords. A number of Semantic Search methods have been suggested in that context.

Natural Language Processing (NLP) techniques have long been applied to better grasp the semantics of the query or documents. Often, Part-Of-Speech (POS) tagging is first applied on the textual content in order to assign grammatical tags (such as *noun*, *conjunction* or *verb*) to individual words. Such assignment is highly accurate for well-formed sentences (Manning 2011), but much more challenging for short texts such as queries (Hua et al 2015). POS tags can then by used to better discriminate textual keywords, for example for Named-Entity Recognition (NER), where the task is to identify which keywords correspond to real-world entities, or for co-reference resolution, where the task is to identify all keywords referring to the same entity in text. Sentence parsing takes NLP analyses to the next level by aiming at capturing the overall structure of sentences, typically through a dependency parse tree.

NLP methods are often combined with lexical resources or third-party sources to retrieve more relevant results. The main idea in this context is to identify entities in the textual query or content and to match them to their counterpart in a third-party resource to improve the search results. Voorhees (1993) proposed an early approach in that sense that leverages WordNet to disambiguate word senses and hence improve search results. Pehcevski et al (2008) analyze the structure of Wikipedia to better rank relevant entities in response to a search request. Kaptein et al (2010) use Wikipedia to better characterize and identify entities when searching for entities in document collections, while Schuhmacher et al (2015) combine different features from the documents, the entity mentions and Wikipedia using a learning-to-rank approach to improve the search results.

Conceptually similar approaches have been proposed in the context of the Semantic Web, by leveraging the structure or contents of a knowledge base to better grasp the context of queries or entities appearing in textual documents. Tran et al (2007), for instance, propose an ontology-based interpretation of natural language queries for Semantic Search. The authors translate a keyword query into a description logic, conjunctive query that can then be evaluated with respect to an underlying knowledge base. Schuhmacher and Ponzetto (2013) exploit entities and semantic relations from the DBpedia knowledge base to cluster the results of a search engine into more meaningful groups. Prokofyev et al (2015) leverage an ontology to better resolve co-references in textual documents for Semantic Search tasks.

Machine Learning techniques are often used for Semantic Search, to power some of the approaches described above, but also to capture the context and semantics of the words or entities appearing in documents. One of the main intuitions in this context is that words that occur in similar contexts are likely to be semantically similar. Early approaches leveraging this observation built high-dimensional matrices capturing the co-occurrence of words in a certain context (e.g., within a window of a few words), and hence the similarity between words (Lund and Burgess 1996). Each word is in that case represented by a sparse vector in a high-dimensional space. Lower-dimensional embeddings can then be created by applying standard matrix factorization techniques like Principal Component Analysis.

More recently, Mikolov et al (2013) suggested a new word embedding technique to generate dense vector representations of words efficiently. The method works by maximizing the co-occurrence probability of words appearing within a certain context window using a relatively simple neural network. This opened the door to numerous applications, by efficiently generating vector representations of words from large text corpora and feeding them into subsequent machine learning models. Approaches to improve Entity Recognition (Siencnik 2015), web search query expansion (Grbovic et al 2015), or web search ranking (Nalisnick et al 2016) have for example been explored in the context of Semantic Search.

Knowledge Base Search

A number of Semantic Search approaches target large and declarative knowledge bases (a.k.a ontologies or knowledge graphs) instead of document collections. Such knowledge bases can be expressed in many different ways that are typically derived from Semantic Web standards such as RDF or OWL. Google's Knowledge Graph, DBpedia (Bizer et al 2009), Yago (Rebele et al 2016) or Wikidata (Vrandecic and Krötzsch 2014) are well-known examples of that trend. Users can express their queries through two main modalities in this case: either as structured (e.g., SPARQL) queries, or as natural language (e.g., keyword) queries.

Horrocks and Tessaris (2002) introduced an early formal approach to answer structured queries posed against ontologies. Their algorithm return sound a complete results to conjunctive queries leveraging reasoning techniques and Description Logics. Stojanovic et al (2003) present a method to rank results in ontology-based search. The authors consider conjunctive queries, and combine logical inference with an analysis of the contents of the knowledge base to retrieve more relevant results. Maedche et al (2003) introduce a meta-ontology and a registry to improve search queries targeting ontologies. Their solution leverages WordNet to match entities appearing in the ontology to lexical entries and to guide the search process.

Pound et al (2010) introduce the ad-hoc object retrieval task for searching for resources (e.g., entities, types or relations) over knowledge bases using natural language queries. The authors also propose a baseline technique for answering such queries based on term frequencies as well as an evaluation methodology. Tonon et al (2012) propose an improved search technique for ad-hoc object retrieval exploiting sequentially an inverted index to answer keyword queries and a graph database to improve the search effectiveness by automatically generating declarative queries over an RDF graph.

Hybrid approaches leveraging both textual and structured contents have also been suggested. Rocha et al (2004), for instance, combine a traditional search engine with graph exploration techniques to answer keyword queries on an ontology. Zhang et al (2005) suggest a new model to search semantic portals, where both documents and structured data are available. Their method is based on creating textual representations for all entities in the structured repository such that they can also be indexed and searched through classical Information Retrieval techniques.

Word Embeddings techniques (see above) have also been adapted to power Semantic Search on knowledge bases. RDF2Vec (Ristoski and Paulheim 2016), for instance, learns vectorial representations of entities in RDF graphs. (Wang et al 2017) provide a survey of embedding approaches for knowledge bases and classify the models into two main families: translation-based techniques, which interpret relations in the knowledge base as a translation vector between the two entities connected by the relation, and semantic matching models, which exploit similarity-based scoring functions to create the embeddings.

Systems

Leading industrial search engines, such as Bing, Yandex or Google, all implement Semantic Search in one way or another, but typically do not describe in detail the techniques they leverage. A number of Semantic Search systems have been described or open-sourced, however, and are summarized below.

TAP (Guha et al 2003) is an early Semantic Search framework. TAP focuses on entity search queries expressed as keywords, and augments traditional results (documents) with semi-structured data returned from a knowledge base. The knowledge base is also used to better filter and sort the list or returned documents in that context.

A number of Semantic Search systems focusing on RDF and ontologies have been proposed. Swoogle (Ding et al 2004) offers semantic search over a knowledge base represented in RDF thanks to an inverted index and a database storing metadata about all entities. SWSE (Hogan et al 2011) follows the typical architecture of a search engine but operates on RDF data also. SWSE results are ranked by running a classical PageRank algorithm on a graph connecting the URIs appearing in the RDF triple to their source on the Web.

SemSearch (Lei et al 2006) is a search engine for the Semantic Web that hides the complexity of the underlying RDF data to the user. SemSearch accepts keyword queries from the user, translates the user queries into formal queries by exploiting the labels of the entities in the knowledge base, runs the resulting query in the knowledge base and finally ranks the results by taking into account the number of keywords the search results satisfy. Sindice (Oren et al 2008) is a Semantic Search engine and look-up service that focuses on scaling to very large quantities of semi-structured data. It supports keyword and URI-based search as well as structured queries.

SHOE (Heflin and Hendler 2000) is an early Semantic Search system collecting semi-structured annotations from the web and storing them in a knowledge base. It offers a GUI to formulate ontology-based structured queries to find webpages. The Watson system (d'Aquin and Motta 2011) works similarly by collecting, analyzing and giving access to ontologies and semantic data available online. It supports both keyword and SPARQL search queries.

SCORE (Sheth et al 2002) is a platform supporting the creation and maintenance of large knowledge bases. It supports ontology-driven Semantic Search capabilities by extracting facts and metadata from web sources via text mining techniques.

Nordlys (Hasibi et al 2017) is an open-source toolkit for Semantic Search. The toolkit supports a number of features including detecting entities in natural language queries, cataloging entities from a knowledge base (by default DBpedia), interlinking entities, and retrieving entities.

Schema.org (Mika 2015), finally, is not a system per se, but rather a standardization effort founded by leading search engines (including Google, Microsoft, Yahoo and Yandex). Its mission is to create and promote schemas to embed semi-structured data in web documents (and beyond) in order to facilitate Semantic Search capabilities online.

Cross-References

Knowledge Graph Embeddings, Reasoning at Scale, Semantic Interlinking.

References

Bast H, Buchhold B, Haussmann E (2016) Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval 10(2-3):119–271, DOI 10.1561/1500000032, URL http://dx.doi.org/10.1561/1500000032

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) Dbpedia - A crystallization point for the web of data. J Web Sem 7(3):154–165, DOI 10.1016/j.websem.2009.07.002, URL https://doi.org/10.1016/j.websem.2009.07.002

d'Aquin M, Motta E (2011) Watson, more than a semantic web search engine. Semant web 2(1):55–63, URL http://dl.acm.org/citation.cfm?id=2019470.2019476

Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J (2004) Swoogle: A search and metadata engine for the semantic web. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM

'04, pp 652–659, DOI 10.1145/1031171.1031289, URL http://doi.acm.org/10.1145/1031171.1031289

Grbovic M, Djuric N, Radosavljevic V, Silvestri F, Bhamidipati N (2015) Context- and content-aware embeddings for query rewriting in sponsored search. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '15, pp 383–392, DOI 10.1145/2766462.2767709, URL http://doi.acm.org/10.1145/2766462.2767709

Guha R, McCool R, Miller E (2003) Semantic search. In: Proceedings of the 12th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '03, pp 700–709, DOI 10.1145/775152.775250, URL http://doi.acm.org/10.1145/775152.775250

Hasibi F, Balog K, Garigliotti D, Zhang S (2017) Nordlys: A toolkit for entity-oriented and semantic search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '17, pp 1289–1292, DOI 10.1145/3077136.3084149, URL http://doi.acm.org/10.1145/3077136.3084149

Heflin J, Hendler J (2000) Searching the web with shoe. In: AAAI Workshop on Artificial Intelligence for Web Search, pp 35–40

Hogan A, Harth A, Umbrich J, Kinsella S, Polleres A, Decker S (2011) Searching and browsing linked data with swse: The semantic web search engine. Web Semantics: Science, Services and Agents on the World Wide Web 9(4):365 – 401, DOI https://doi.org/10.1016/j.websem.2011.06.004, URL http://www.sciencedirect.com/science/article/pii/S1570826811000473, jWS special issue on Semantic Search

Horrocks I, Tessaris S (2002) Querying the semantic web: A formal approach. In: Horrocks I, Hendler J (eds) The Semantic Web — ISWC 2002, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 177–191

Hua W, Wang Z, Wang H, Zheng K, Zhou X (2015) Short text understanding through lexical-semantic analysis. In: 2015 IEEE 31st International Conference on Data Engineering, pp 495–506, DOI 10.1109/ICDE.2015.7113309

Kaptein R, Serdyukov P, de Vries AP, Kamps J (2010) Entity ranking using wikipedia as a pivot. In: Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010, pp 69–78, DOI 10.1145/1871437.1871451, URL http://doi.acm.org/10.1145/1871437.1871451

Lei Y, Uren VS, Motta E (2006) Semsearch: A search engine for the semantic web. In: Managing Knowledge in a World of Networks, 15th International Conference, EKAW 2006, Podebrady, Czech Republic, October 2-6, 2006, Proceedings, pp 238–245, DOI 10.1007/11891451_22, URL https://doi.org/10.1007/11891451_22

Lund K, Burgess C (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers 28(2):203–208, DOI 10.3758/BF03204766, URL https://doi.org/10.3758/BF03204766

Madhu G, Govardhan A, Rajinikanth TV (2011) Intelligent semantic web search engines: A brief survey. CoRR abs/1102.0831, URL http://arxiv.org/abs/1102.0831, 1102.0831

Maedche A, Motik B, Stojanovic L, Studer R, Volz R (2003) An infrastructure for searching, reusing and evolving distributed ontologies. In: Proceedings of the 12th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '03, pp 439–448, DOI 10.1145/775152.775215, URL http://doi.acm.org/10.1145/775152.775215

Mäkelä E (2005) Survey of semantic search research. URL https://seco.cs.aalto.fi/publications/2005/makela-semantic-search-2005.pdf

Mangold C (2007) A survey and classification of semantic search approaches. Int J Metadata Semant Ontologies 2(1):23–34, DOI 10.1504/IJMSO.2007.015073, URL http://dx.doi.org/10.1504/IJMSO.2007.015073

Manning CD (2011) Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In: Gelbukh AF (ed) Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 171–189

Mika P (2015) On schema.org and why it matters for the web. IEEE Internet Computing 19(4):52–55, DOI 10.1109/MIC.2015.81

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. CoRR abs/1301.3781, URL http://arxiv.org/abs/1301.3781, 1301.3781

Nalisnick E, Mitra B, Craswell N, Caruana R (2016) Improving document ranking with dual word embeddings. In: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, WWW '16 Companion, pp 83–84, DOI 10.1145/2872518.2889361, URL https://doi.org/10.1145/2872518.2889361

Oren E, Delbru R, Catasta M, Cyganiak R, Stenzhorn H, Tummarello G (2008) Sindice.com: a document-oriented lookup index for open linked data. IJMSO 3(1):37–52, DOI 10.1504/IJMSO.2008.021204, URL https://doi.org/10.1504/IJMSO.2008.021204

Pehcevski J, Vercoustre AM, Thom JA (2008) Exploiting locality of wikipedia links in entity ranking. In: Macdonald C, Ounis I, Plachouras V, Ruthven I, White RW (eds) Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 258–269

Pound J, Mika P, Zaragoza H (2010) Ad-hoc object retrieval in the web of data. In: Proceedings of the 19th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '10, pp 771–780, DOI 10.1145/1772690.1772769, URL http://doi.acm.org/10.1145/1772690.1772769

Prokofyev R, Tonon A, Luggen M, Vouilloz L, Difallah DE, Cudré-Mauroux P (2015) Sanaphor: Ontology-based coreference resolution. In: Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366, Springer-Verlag, Berlin, Heidelberg, pp 458–473, DOI 10.1007/978-3-319-25007-6_27, URL https://doi.org/10.1007/978-3-319-25007-6_27

Rebele T, Suchanek FM, Hoffart J, Biega J, Kuzey E, Weikum G (2016) YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In: The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II, pp 177–185, DOI 10.1007/978-3-319-46547-0_19, URL https://doi.org/10.1007/978-3-319-46547-0_19

Ristoski P, Paulheim H (2016) Rdf2vec: Rdf graph embeddings for data mining. In: Groth P, Simperl E, Gray A, Sabou M, Krötzsch M, Lecue F, Flöck F, Gil Y (eds) The Semantic Web – ISWC 2016, Springer International Publishing, Cham, pp 498–514

Rocha C, Schwabe D, Aragao MP (2004) A hybrid approach for searching in the semantic web. In: Proceedings of the 13th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '04, pp 374–383, DOI 10.1145/988672.988723, URL http://doi.acm.org/10.1145/988672.988723

Schuhmacher M, Ponzetto SP (2013) Exploiting dbpedia for web search results clustering. In: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, ACM, New York, NY, USA, AKBC '13, pp 91–96, DOI 10.1145/2509558.2509574, URL http://doi.acm.org/10.1145/2509558.2509574

Schuhmacher M, Dietz L, Paolo Ponzetto S (2015) Ranking entities for web queries through text and knowledge. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '15, pp 1461–1470, DOI 10.1145/2806416.2806480, URL http://doi.acm.org/10.1145/2806416.2806480

Sheth A, Bertram C, Avant D, Hammond B, Kochut K, Warke Y (2002) Managing semantic content for the web. IEEE Internet Computing 6(4):80–87, DOI 10.1109/MIC.2002.1020330

Siencnik SK (2015) Adapting word2vec to named entity recognition. In: Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Institute of the Lithuanian Language, Vilnius, Lithuania, pp 239–243, URL http://aclweb.org/anthology/W/W15/W15-1830.pdf

Stojanovic N, Studer R, Stojanovic L (2003) An approach for the ranking of query results in the semantic web. In: Fensel D, Sycara K, Mylopoulos J (eds) The Semantic Web - ISWC 2003, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 500–516

Tonon A, Demartini G, Cudré-Mauroux P (2012) Combining inverted indices and structured search for ad-hoc object retrieval. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '12, pp 125–134, DOI 10.1145/2348283.2348304, URL http://doi.acm.org/10.1145/2348283.2348304

Tran T, Cimiano P, Rudolph S, Studer R (2007) Ontology-based interpretation of keywords for semantic search. In: Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, Springer-Verlag, Berlin, Heidelberg, ISWC'07/ASWC'07, pp 523–536, URL http://dl.acm.org/citation.cfm?id=1785162.1785201

Voorhees EM (1993) Using wordnet to disambiguate word senses for text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '93, pp 171–180, DOI 10.1145/160688.160715, URL http://doi.acm.org/10.1145/160688.160715

Vrandecic D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Commun ACM 57(10):78–85, DOI 10.1145/2629489, URL http://doi.acm.org/10.1145/2629489

Wang Q, Mao Z, Wang B, Guo L (2017) Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering 29(12):2724–2743, DOI 10.1109/TKDE.2017.2754499

Zhang L, Yu Y, Zhou J, Lin C, Yang Y (2005) An enhanced model for searching in semantic portals. In: Proceedings of the 14th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '05, pp 453–462, DOI 10.1145/1060745.1060812, URL http://doi.acm.org/10.1145/1060745.1060812