Document Relation System Based on Ontologies for the Security Domain

Janine Hellriegel[1], Hans Georg Ziegler[1], and Ulrich Meissen[1]

[1] Fraunhofer Institute for Open Communication Systems (FOKUS),
Kaiserin Augusta Allee 31, 10589 Berlin, Germany

{ Janine.Hellriegel, Hans.Georg.Ziegler,
Ulrich.Meissen }@fokus.fraunhofer.de

Abstract. Finding semantic similarity or semantic relatedness between unstructured text documents is an ongoing research field in the semantic web area. For larger text corpuses often lexical matching – the matching of shared terms – is applied. Related sematic terms and concepts are not considered in this solution. Also documents that use heterogeneous perspectives on a domain could not be set into a relation properly. In this paper, we present our ongoing work on a flexible and expandable system that handles text documents with different points of view, languages and level of detail. The system is presented in the security domain but could be adapted to other domains. The decision making process is transparent and the result is a ranked list.
Keywords: Document Relation, Security, Ontology, Semantic Relatedness

1 Introduction

The amount of available information in the Internet is growing day by day. It is difficult to keep an overview of relevant data in a domain, especially if different kinds of views on the same topic are considered. An expert is using different words and level of detail in contrast to a normal user, but they describe exactly the same concept. Having a database consisting of documents authored from people with different levels of expertise, language skills and ambitions imposes a big challenge on a semantic search algorithm. The usage of long texts as search input enables a wider range of search terms, which is the foundation to detect a larger spectrum of documents. The relevant results are documents related to the input query text document. A basic method to compare two text documents is the vector space model [2], which relates the text similarity to the amount of similar words. However, semantically related words are not considered. Knowledge based similarity measures use lager document corpuses and external networks like WordNet or Wikipedia to analyze co-occurrences and relations. An overview of theses techniques is presented in [3] but most of the methods just work for a couple words as search query. Although all documents affiliate to one domain (e.g. the security domain) lexical matching and knowledge-based measure don't retrieve a sufficient number of related documents. Another measure, the

Ontology based matching includes concepts and heterogeneous relations. Wang [7] proposes a system to relate documents using the concepts found in WordNet. But the measurement step still depends on words and heterogeneous concepts could not be related. In the security and safety domain only specialized ontologies exist [5], [6], that mainly focus on the security of information systems. An attempt to combine different ontologies was made by [1] but could not express the diversity of the domain also addressing e.g. security of citizens, infrastructures or utilities. As the mentioned references show, a system that searches for related text documents in a clear and traceable way is not yet developed. At the moment no ontology exists that would match the terminology of the whole security domain. Therefore a new, more general, ontology as well as a general system are developed.

2 System of Semantic Related Documents

The fundament to measure semantic relatedness between two documents are terms. A terminology is built, which is used to compare all documents quickly and determine their relations. The whole system is divided into three steps. Figure 1 displays an overview of the whole system.

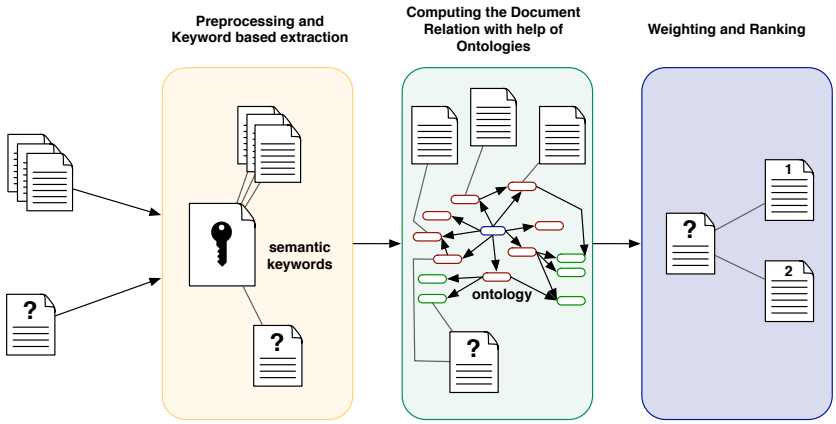

Fig. 1. System overview with three steps to determine document relatedness

A possible scenario is the goal to find related work and potential partners for a project idea. In the first step predefined keywords are extracted from the project descriptions and organization profiles as well as the query text containing the project idea to discover terms that characterize the documents. Each document is now represented with the detected keywords. In the second step the text documents are classified in the ontology according to their keywords in order to discover further relations. If the keyword *maritime borders* appears in the query document, all relations from this keyword to others, like *border surveillance*, are used. The ontology helps to discover related keywords and therefore related documents. With the help of a weighting algorithm a ranked list of related documents is the result in the last step.

2.1 Preprocessing and Keyword based extraction

In order to extract the valuable terms from the documents, a manually created keyword list is used and their term frequency for each document is determined. Comparing the occurrence of the keywords gives a first measure for the relatedness. The more terms the texts have in common, the more related they are. However, different views and special relations are not yet taken into account. In order to extract the keywords all documents are preprocessed with a tokenization on term bases. Further, stemming algorithms are used to transform all terms in the documents as well as all keywords to their base form. The keyword list was developed by early-warning system experts together with civil protection and police specialists. It contains about 500 English words relevant to the security domain but still could be modified or extended. Automatic keyword extraction algorithms are not suitable since they produce too much noise and could not hold up to the quality of the keywords list. All keywords can be translated semi-automatically. Therefore the system supports different languages. Synonyms, categories and other semantic relevant words are added by using BabelNet [4]. From the term *video surveillance* the terms *surveillance camera, cctv, video home security system* are derived. In total a keyword list with over 4000 terms has been produced. In this way it is ensured, that only domain related terms are found.

2.2 Computing the Document Relation with help of Ontologies

In the case when related documents don't contain identical or similar terms, an ontology or terminological net can be used in order to improve the calculation of the relatedness. The relation between a technical and a user view could only be determined over a shared concept. Using the heterogeneous paths between the terms in the graph-based knowledge representation, new relations between the documents are revealed. Not only the distances in the terminological net are considered, also the type of relation like *is-a* or *part-of* between the terms determines the relatedness of the text documents. In this way, for each detected keyword in the query document, related keywords could be found. Texts containing the related keywords are most likely to correlate with the query document. A new ontology in the security domain is manually built at the moment, containing the original 500 keywords, relations from BabelNet and a taxonomy created by security researchers. The taxonomy is loosely based on a project categorization for the recent FP 7 Cordis security call [8].

2.3 Weighting and Ranking

A ranked list of texts related to the query document is the result of the system. Two measurements are used to rank the results, first is the weighting of the original keywords and second is the type of relation between the keywords. Not all retrieved terms are equally important to distinguish the texts. The term *security* is important but very general and can be found in a lot of documents. Due to the low entropy of the term, it does not help to find unique relations. In contrast, the term *body scanner* is more useful to find related documents. A term weighting is applied with the tf-idf

statistic [2] to identify significant terms. As document corpus the FP 7 Cordis security call project descriptions are used. Secondly the relation between two specific keywords (*body scanner* and *metal detector*) is ranked higher then a relation between a specific keyword and a more general keyword (*body scanner* and *airport security*).

3 Conclusion and Future Work

With the presented system, a ranked list of related documents can be retrieved. Regardless what kind of view or level of detail they contain. The system describes a general sequence of functions and could be adapted to other domains if a corresponding keyword list and ontology are available. In the music domain e.g. artist profiles could be related to genre or instrument descriptions. The system is based on a simple method but achieves good results because it works close to the domain. In addition, it allows the evaluation of the results and to understand why documents are identified as related. The system is still work in progress, the next steps are to complete the development of the ontology and to evaluate the chosen keywords. Further evaluations concerning the accuracy as well as user satisfaction have to be performed.

References

1. Liu, Shuangyan, Duncan Shaw, and Christopher Brewster: Ontologies for Crisis Management: a Review of State of the Art in Ontology Design and Usability. In: Proceedings of the Information Systems for Crisis Response and Management Conference (2013)
2. Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze: Introduction to Information Retrieval. Vol. 1. Cambridge university press Cambridge (2008)
3. Mihalcea, Rada, Courtney Corley, and Carlo Strapparava: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: AAAI, 6:775–80 (2006)
4. Navigli, Roberto, and Simone Paolo Ponzetto: BabelNet: Building a Very Large Multilingual Semantic Network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 216–25. Association for Computational Linguistics (2010)
5. Ramanauskaite, Simona, Dmitrij Olifer, Nikolaj Goranin, and Antanas Čenys: Security Ontology for Adaptive Mapping of Security Standards. In: International Journal of Computers Communications & Control 8, no. 6 (2013)
6. Souag, Amina, Camille Salinesi, and Isabelle Comyn-Wattiau: Ontologies for Security Requirements: A Literature Survey and Classification. In: Advanced Information Systems Engineering Workshops, 61–69. Springer (2012)
7. Wang, James Z., and William Taylor: Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method. In: Web Intelligence, IEEE/WIC/ACM International Conference On, 395–401. IEEE (2007)
8. FP7 Cordis Project, http://cordis.europa.eu/fp7/security/home_en.html

Acknowledgement

This work has received funding from the Federal Ministry of Education and Research for the security research project “fit4sec” under grant agreement no. 13N12809.