Semantic Web Technologies to Improve Customer Service*

Kerstin Denecke
Forschungszentrum L3S
Appelstr. 9a
30167 Hannover, Germany
denecke@l3s.de

Wladimir Krasnov
Forschungszentrum L3S
Appelstr. 9a
30167 Hannover, Germany
krasnov@l3s.de

Gideon Zenz
Forschungszentrum L3S
Appelstr. 9a
30167 Hannover, Germany
zenz@l3s.de

ABSTRACT

In this paper, we present an approach that exploits semantic web technologies to categorize specialized text and to create hierarchical facets representing the document content. For this purpose, domain knowledge represented by a thesaurus with relevant, domain-specific terms is used to identify relevant terms. Based on dependency information between single terms provided by the thesaurus (hypernomy, hyponymy), we create hierarchical facets representing the content of the text. The algorithm is applied to a collection of service messages and shows promising results in text categorization.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Hierarchical Facets, Text Classification, Search

1. INTRODUCTION

In today's information society, quick access to relevant information is essential for individuals, but also for companies. For example, it is crucial for service departments to find relevant information that helps to answer customer requests in time to avoid long process durations for the customer. But, relevant knowledge is often stored in non-retrievable form, which causes huge difficulties in knowledge discovery for employees. Further, a long learning period is needed for new employees to go into the depth with existing (company internal) knowledge and to learn how to formulate the "right" request for finding relevant information. Even with existing retrieval systems, formulating the "right" request that results in an appropriate number of suited results is difficult, as there are many different ways to describe a problem with natural language.

Service messages are documented in natural language and are therefore stored in unstandardized, unstructured manner. To allow efficient access and reuse of this data, we describe an approach to store service messages in a standardized way and to structure the content of a text collection hierarchically into facets. By visualizing and navigating this hierarchy, a user is able to specify constraints on the items selected from the repository. The facets help to discover similar service messages even if they use different terms to describe similar issues. Furthermore, the service message database can be browsed in an intuitive manner.

Currently, two methods are quite popular to group search results appropriately which is clustering and faceted categorization. Clustering groups items according to some measure of similarity. The Latent Dirichlet Allocation (LDA, [1]) algorithm generates a probabilistic model to describe the content of texts and clusters them based on this model. Scatter/Gather offers a navigation system based on document clustering [3].

In faceted classification, a set of category hierarchies is built, rather than a single large category hierarchy [5]. These capture the different facets, i.e., dimensions or features, relevant to a collection. In [4], an unsupervised approach for facet extraction is presented relying upon WordNet and Wikipedia, to identify useful facet terms.

The approach presented in this paper can mainly be seen as a modification and extension of the Castanet algorithm presented by Stoica et al. [6]. This algorithm generates automatically hierarchical faceted metadata from text based on WordNet and WordNet Domains. We modified and adapted the original algorithm to fit the given scenario. In particular, this algorithm is customized to the domain of mechanical engineering by using a domain specific thesaurus. Further, it is extended by a more sophisticated text preprocessing comprising stemming of words and resolution of compound nouns. The Castanet approach only considers nouns to generate the structure while our extension considers all relevant words. Finally, the algorithm is applied and tested on a collection of service messages of the engineering domain.

2. METHOD

Similar to the CastaNet Algorithm, our method requires a lexical knowledge base. Since our system targets at processing documents of the engineering domain, a domain-unspecific lexical resource such as WordNet[1] or GermaNet[2] is unsuited since relevant domain-specific terms are missing. Therefore, we decided to base our algorithm to the FIZ Thesaurus[3].

The FIZ thesaurus provides 58,300 technical terms in German and 63,500 technical terms in English which are related through links indicating synonymy, hierarchy and semantic relation. The thesaurus covers the vocabulary of different

*This work is funded by AiF under 15452 N

[1]http://wordnet.princeton.edu/
[2]http://www.sfs.uni-tuebingen.de/GermaNet/
[3]http://www.fiz-technik.de/fiz/thesaurus.htm

engineering subfields and was originally created to extend online literature repositories and to improve the document retrieval.

For our algorithms, the thesaurus is stored into RDF which allows for its manual extension and its efficient use for creating hierarchies. Each word is represented by a node with a unique id, the actual word and the stemmed version of the word. Nodes can be connected as synonyms, parents, or other semantic relations. A term A is considered as parent of term B if A is a generic term to B.

To represent a document through categories of the underlying thesaurus the following processing steps are conducted. First, the document is preprocessed, i.e. special characters and other punctuation marks are removed and the words are split into their linguistic segments. This is necessary to be able to identify compound nouns, which occur very often in service messages due to the reduced length of these messages and the compact language. In addition, the words are stemmed using the Stemmer presented in [2] and looked up in the FIZ thesaurus. The matched words are assembled in a domain specific document term list which in turn is extended by synonyms of the terms that are collected using the synonym relationship provided by the thesaurus.

Finally, a directed graph is constructed where the terms of the expanded term list are the leaves. Starting from the term, the hyperonomy relations provided by the thesaurus are iteratively used for constructing a hierarchical tree. In more detail, for each term (leaf) parent nodes are collected from the thesaurus and are inserted into the graph. The result is a connected graph that resembles a hierarchical semantic representation of the document.

Furthermore, all paths are attributed with a count that specifies the usage frequency of each concept. This frequency information is used to select facets as categories for the document under consideration. In particular, the most probable generic concepts are collected for categorizing the document.

3. EVALUATION

The introduced method is evaluated on service messages of the mechanical engineering domain. The collection consists of 4,884 documents written in German. In average, each message comprises 20-30 words and are combined to incomplete sentences where a hotline employee summarizes the error description or request of a customer. From this collection, 200 service messages are randomly selected. Four persons were involved in the evaluation. For each test document, the evaluating person was confronted with the system generated hierarchical graph and had to select the best suited categories describing the document from this graph. The categories chosen by the algorithm remained hidden to the evaluators. In particular, the evaluation examines whether the system assigns the test documents to the same categories as the evaluators.

The system achieved a precision of 0.93 and a recall of 0.86. Furthermore, the time to categorize manually was measured. The evaluators needed in average 2.5 minutes to select relevant categories, obviously due to the complexity of the domain and the shortness of the service message. Compared to this, the proposed method takes only 0.5 seconds for classifying a service message. The hierarchical graphs for the complete data set were generated in 130 seconds on a 1.6 GHz Pentium Processor with 2 GB RAM. Since the presented system performs very well in terms of accuracy and time, we can conclude, that this method can help to reduce the time for categorization significantly.

Errors occur when abbreviations were used in a text or terms could not be found in the lexical database. A manual or semi-automatically extension of the lexical resource with relevant abbreviations could help to improve the system's accuracy. The system also fails when confronted with terms with writing errors. Technologies for error correction or soundex- or metaphone technologies[4] could help to deal with this problem.

In contrast to existing algorithms that allow for the generation of hierarchical facets for general texts [6], we showed how domain-specific knowledge can be exploited efficiently to create a faceted representation of technical texts. Our representation contains only technical terms which is crucial for document retrieval purposes. Through the preprocessing of the natural language text messages by means of stemming and analysis of compound nouns, the system is able to identify morphological variants and to identify terms even if they are hidden within compound nouns.

We tested the algorithm on German texts only since similar texts in English were unavailable in this project. Since the FIZ thesaurus already contains terms in English, the system can be applied to this language, too. An adaption of the linguistic preprocessing is required.

4. CONCLUSIONS

In this paper, an approach to generate hierarchical facets to highly specialized texts of the engineering domain was introduced. We showed that domain-specific knowledge can be successfully used to generate such facet representation. By exchanging the domain knowledge, the approach can be transferred to other domains. A hierarchically structured domain knowledge is for example available in the biomedical domain through the UMLS. In future work, we will study the applicability of the proposed approach to this domain. Additionally, we will study how the generated facets can improve document retrieval.

5. REFERENCES

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *JMLR*, 3, 2003.

[2] Jörg Caumanns. A fast and simple stemming algorithm for german words. Technical report, FU Berlin, 1999.

[3] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92*, pages 318–329, New York, NY, USA, 1992. ACM.

[4] Wisam Dakka and Panagiotis G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *ICDE '08*, pages 466–475, Washington, DC, USA, 2008. IEEE Computer Society.

[5] Marti A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, 2006.

[6] Emilia Stoica, Marti Hearst, and Megan Richardson. Automating creation of hierarchical faceted metadata structures. In *Proc. NAACL-HLT 2007*, pages 244–251, 2007.

[4] http://www.sound-ex.de