Text-based Knowledge Acquisition for Ontology Engineering

Uwe Quasthoff
NLP Group, CS Inst., Leipzig University
Augustusplatz 10/11
Leipzig, Germany, D-04109
quasthoff@informatik.uni-leipzig.de

Christian Wolff
NLP Group, CS Inst., Leipzig University
Augustusplatz 10/11
Leipzig, Germany, D-04109
wolff@informatik.uni-leipzig.de

Abstract

This paper describes an approach towards ontology engineering that makes use of text technology for extracting relevant semantic relations from document collections. A short description of corpus characteristics and examples of statistical text analysis results show how input for ontology design can be generated automatically. The Topic Map standard is used as an example for standardised representation of ontologies and a toolset for generating raw Topic Maps is described. Finally practical applications of this approach are described and areas of future research in the refinement of ontology generation are described.

Introduction

There is both a theoretical challenge and a practical demand for describing the relation between given concepts occurring in arbitrary text collections. Some of the constituents of the solution are quite clear: First, text analysis with statistical methods and, especially, efficiently implemented collocation measures can be used to find connected concepts. Thus, relevant associations between concepts can be identified for any text corpus while the semantic nature or type of these associations remains. Second, a set of relevant relations can easily be identified. Finally, the problem of representation can be solved using a recently developed standard for information structuring like the ISO standard for Topic Maps in its XML version. For the remaining problem of naming relations, a combination of pattern matching and machine learning with human assistance is proposed.

This paper is organized as follows: Chapter 1 discusses a text-based approach towards ontology engineering; ch. 2 gives a short introduction on our text mining approach for corpus analysis along with some examples of analysis results and visualizations. Ch. 3 describes automatic Topic Map generation as a follow-up process to our text mining analysis. Finally, practical applications of this approach are briefly mentioned and future research areas especially concerning more refined Topic Map input are briefly introduced.

1 Ontology Engineering and Text Mining

“Ontology engineering has as its goal effective support of ontology development throughout its life cycle – design, evaluation, maintenance, deployment, mapping, integration, sharing, and reuse”. [Gruninger & Lee 02:40]. Given this definition, our text mining-based approach is directed at given empirically derived input for ontology design. While in traditional approaches software tools simply provide *support* for the knowledge worker in modeling relevant knowledge (ontology editors, data modeling tools etc., cf. Kim 02:50f]), our approach will generate relevant material input for ontology design by extracting relevant relations from text collections which represent a domain for which an ontology is to be defined.

The general idea behind this approach is that unstructured text is the most common representation for information in various domains. Starting from the observation that well-known measures for significant collocations can be used to extract relevant relations between concepts from text, we maintain that the general *notion of relatedness* which is the result of such statistical analysis is adequate as input information for ontology design.

2 Text Mining Analysis of Large Corpora

Traditionally, the semantic analysis of a given domain starts with intellectual effort in knowledge identification and acquisition, e. g. by analyzing document indexes containing relevant concepts. To find relations between concepts, one has to look for nearby citations (for instance, in the same chapter) or one has to read and to understand the relevant section. Understanding is, moreover, necessary to classify the relations according to a given scheme.

The aim of *text mining* is the automated extraction of information from large text corpora. This includes collocation analysis: Statistical methods are applied to identify words which occur significantly often together within a predefined text window. Interesting text windows are immediate next neighbors, sentences, and documents. The following section briefly describes our text mining infrastructure and gives some examples of automatically generated results.

2.1 An Infrastructure for Text Corpus Analysis

Since 1994 we have been setting up an infrastructure for processing and analyzing electronic text corpora [see Quasthoff & Wolff 00]. It is available on the web (see http://wortschatz.uni-leipzig.de and http://texttech.de) and comprises one of the largest online corpora for German, English, Italian, and other European languages. It not only offers basic information on words and concepts like frequency and basic morphological and grammatical properties but also semantic information like synonyms, significant collocations ("semantic associations"). At the core of this infrastructures are statistical algorithms for collocation analysis which compute significant collocations for all word types in the corpus using a metric comparable to the log-likelihood measure (for an overview of collocations metrics see [Lemnitzer 98], [Krenn 00], for details on the algorithms used here, see [Heyer et al. 01]).

Type of data	*# of entries*
Word forms	> 6 Million
Sentences	> 25 Million
Grammar	2.772.369
Pragmatics	33.948
Descriptions	135.914
Morphology	3.189.365
Subject areas	1.415.752
Relations	449.619
Collocations	> 8 Million
Index	> 35 Million

Table 1: German Corpus Overview

	German	*English*	*Italian*	*Dutch*	*French*
Tokens	300 m	250 m	143 m	22 m	15 m
Sentences	13.4 m	13 m	9 m	1.5 m	860,000
Types	6 m	1.2 m	870.000	600,000	230,000

Table 2: Basic Characteristics of the Corpora

Tables 1 and 2 give an overview of size and contents of our reference corpora and the contents of the German reference corpus, respectively.

Word: (Wort_nr: 424): System
Frequency Class: 7 (absolute count: 42215)
Subject Area: Allgemeines/Interdisziplinäre Allgemeinwörter (Allgemeines Interdisziplinäre Allgemeinwörter -> REST) Medizin (Medizin)
Morphology: syst|em (=syst_em)
Base Form: System [42215]
Relations and Links with other words:

- *Synonyms:* Aufbau, Gliederung, Lehrgebäude, Ordnungsprinzip, Regierungsform, Staatsform
- *Synonym of:* Arbeitsweise, Bau, Gedankengefolge, Methode, Netz, Netz, Netzwerk, […]
- *Antonyms*: Nichtsystem [1]
- *Other references:* Aufbau, Methode
- *Positive Connotation*: Originalsystem [30], Spezialsystem [4], Profisystem [3], Vollsystem [2], Sondersystem [2], Hauptsystem [1], Supersystem [1]
- *Negative Connotation*: Scheißsystem [2]
- *Part of Phrase:* Duale System [339], Duales System Deutschland [112], Global Positioning System [111], Duale System Deutschland [109], […]
- *Narrower Terms*: SAP-System [2153], Betriebssystem [1553], Immunsystem [1478], Informationssystem [1105], Steuersystem [701], Nervensystem [546], Ökosystem [524], Schulsystem [495], Standardsystem [483], Zielsystem [469], Sozialsystem [433], […]
- *Parts in the same Complex*: Systems [7016], Systeme [5194], Systemen [2284], Systematische [2174], Systemeinstellungen [736], Systemstatus […]

Figure 1 Basic Statistical Data and Declarative Knowledge for System (German Corpus)

The German corpus is currently the largest, especially in terms of additional declarative knowledge such as subject fields which has been collected from various sources instead of being generated automatically by tokenization (like inflected word types) or statistical analysis (like collocations). Figure 1 contains data from the sample entries for *System* (German Corpus) while figure 2 lists the top strongest sentence, left neighbor, and right neighbor collocations for *System*. Please note that due to space restrictions, the examples are heavily abridged; complete examples may be generated using the online access mentioned above.

Top 20 Significant Collocations of *System*: automatisch (3845), Wenn (2413), Daten (1808), ermittelt (1648), prüft (1599), R (1159), erzeugt (1074), zeigt (871), Ihnen (838), Dualen (781), erstellt (761), verwendet (651), bucht (627), ob (568), Ihrem (546), eingeben (530), z.B. (530), Benutzer (487), bzw (478), Feld (472), […]

Significant Left Neighbour Collocations of *System:* Dualen (1017), Ihrem (675), dieses (665), duale (495), Management (411), dualen (372), Dieses (329), politische (253), externen (246), zentralen (214), neues (202), solches (195), ganze (172), politisches (170), ausgeklügeltes (165), logisches (163), Ihr (141), geschlossenes (138), politischen (128), bisherige (123),

Top 20 Significant Right Neighbour Collocations of *System*: automatisch (1663), zeigt (986), ermittelt (790), prüft (731), erzeugt (471), R (360), verzweigt (320), erstellt (316), bucht (277), blendet (242), verwendet (238), schlägt (234), übernimmt (225), berechnet (223), FI-AA (206), hinterlegt (198), löscht (183), vorhanden (173), führt (167), ordnet (163), […]

Figure 2: Different Types of Collocation Sets for System *(German Corpus)*

These sample data are available for any token found in a given corpus and may be accessed either via a Web front-end or by directly accessing the corpus database (application programs) and are the starting point for ontology engineering (see ch. 3 below).

For data visualization the strongest collocations form the set of collocations given for a specific term are selected, those which do not themselves have some interrelationship are filtered out, and a graph drawing algorithm based on simulated annealing is applied to the remaining set [cf. Davidson & Harel 96]. Fig. 3 shows this type of visualization for *System* (German), *system* (English), and *sistema* (Italian), taken from the respective corpora. The graph visualization is available online as well. The technical architecture of our toolset is not discussed here; for detailed information, see [Böhm et al. 02].

2.2 Postprocessing of Collocations

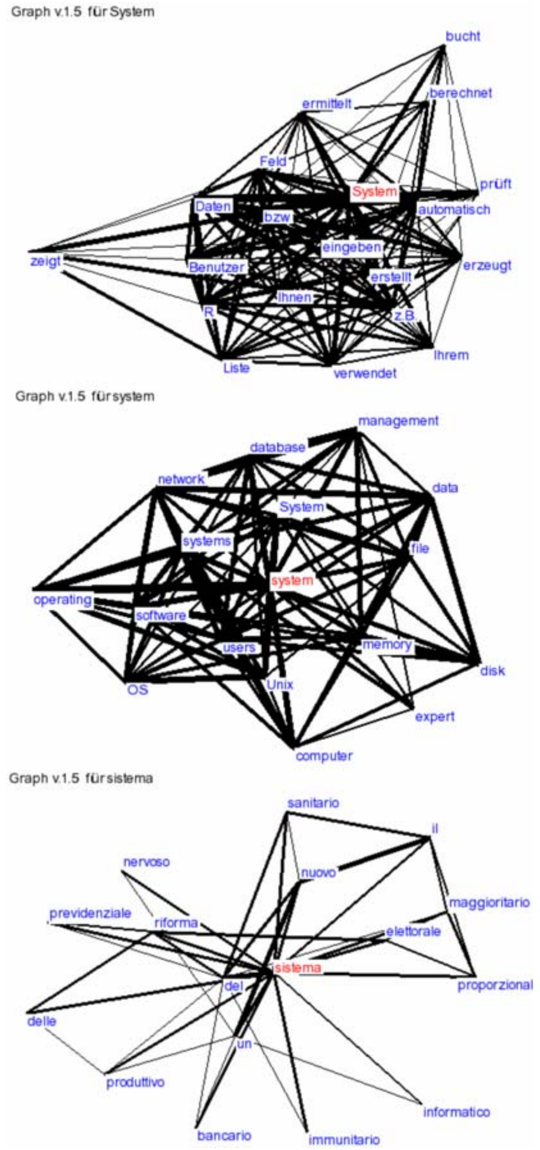


Figure 3: Visualisation of Collocation Sets for System *in the German (*System*), English (*system*), and Italian Corpus (*sistema*)*

A closer look at the collocation sets automatically generated by our text mining engine quickly reveals that they represent various types of semantic relations which are not recognized by the statistical base process. The kind of relation represented by a collocation can often be conjectured using heuristics, linguistic knowledge, or pattern matchers. Some possible approaches are (see [Heyer, Quasthoff, Wolff 02] for a detailed discussion):

1. Analysis of the *numeric value* of our collocation measure.
2. *Positional information* and typical *patterns* for well-defined types of information like person names and titles or company and product names.
3. *Additional knowledge* given by *categorical* and *part-of-speech* information.
4. *Additional knowledge* about *categories* of named entities or given by *subject area codes.*
5. Domain specific collocation sets from another corpus (from a different domain, time, or text type).
6. The application of set operations on collocation sets (intersection, union).

The generation of collocation sets should be interpreted as a kind of basic technology for which various types of applications exist; at the same time it is open for the application of different types of representation standards. In the discussion below, the Topic Map standard is used as representation syntax.

3 Ontology Engineering with Topic Maps

Recently, several standards for information and knowledge structuring and description have emerged, many of them based on or derived from the XML / SGML family of standards for information structuring [see Noy et al. 01:61]. Table 3 lists some major standardization efforts:

Standard	*Focus / Application*
XML (eXtensible Markup Language)	Document structure (mainly syntactic aspect)
RDF (Resource Description Framework / Schema Language)	Description of Metadata, especially for Web Resources; various description schemes or ontologies may be created (RDF Schema Language)
DAML+ OIL (Darpa Agent Markup Language / OIL (Ontology Inference Layer)	Description-logic based standard for ontology engineering
Topic Maps (ISO/IEC 13250 standard	Generic standard for document annotation

Table 3: Major Information Structuring Standards

While these different standardization efforts stem from different scientific and industrial communities, they share a common goal: Simplifying information structuring, access und processing by adding structured metadata to information resources.

3.1 Topic Maps

A *Topic Map* is an information structure to be used as descriptive metadata for arbitrary types of data with document annotation being the most prominent application. *Topic Maps* consist of one or more *Topics*, identified by topic names, describing the resource to which it is attached. Additionally, *topic occurrence* descriptions allow for contextualization of this metadata information. The interrelationship between different topics is formalized by *topic associations* which represent typical semantic relations like *part_of*, *is_a* or *hypero*- and *hyponymy* (for a general introduction to Topic Maps, see [Gerick 00], [Biezunski & Newcomb 01]). Derived from the ISO/IEC HyTime standard for coding multimedia information, Topic Maps are standardized using SGML/XML syntax (XTM standard).

3.2 Topic Map Generation

The generation of Topic Maps and their practical application to information resources involves a great deal of knowledge engineering as the relevant domain has to be intellectually analyzed prior to topic definition and resource description. This phenomenon is common to both, Topic Maps as well as the Semantic Web initiative. Recently, various tools and methods have been developed in order to streamline this process. Some examples shall be mentioned here:

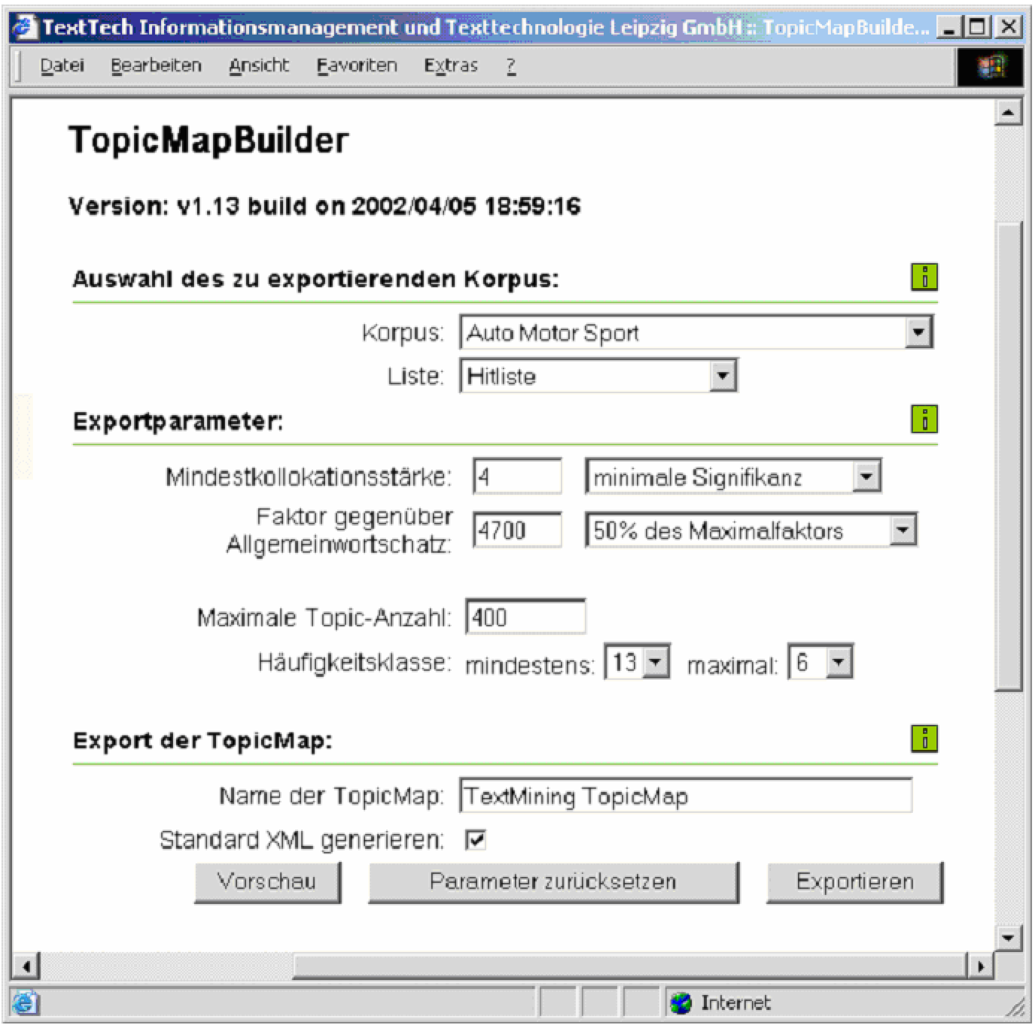
- [Noy et al. 01] describe *Protége-2000*, a modeling tool generating RDF as well as DAML+OIL classes.
- [Maedche & Staab 01] discuss *OntoEdit*, their ontology learning infrastructure which combines various analysis algorithms like text analysis, importing electronic dictionaries, and knowledge databases.
- [Zhou, Booker & Zhang 02] present the *Rapid Ontology Development Method* (ROD) which likewise combines text analysis and relation extraction with domain analysis based on declarative knowledge.

In comparison, our approach is confined to generating raw input for Topic Map definition from the automatic analysis of large text corpora. As should appear to be obvious from the examples given in ch. 2, the results of our collocation analysis yield appropriate material for defining semantic relations. As this analysis is computed for every word type in a corpus, it is not well suited for selecting *central* topics for a certain information collection. In order to achieve this, we employ two different strategies which may be combined and enhanced by additional intellectual revision steps:

- *Topic Map bootstrapping*, and
- *Topic Map optimization* for a given Topic Map which is iteratively refined in this process.

In the first case topic candidates are generated by running a comparative analysis of a domain-specific corpus against a (much larger) reference corpus. Significant concepts are filtered out and used as starting points for Topic Map generation. This process can be controlled using the *TopicMapBuilder*, a tool with a web-based interface for fine-tuning generation parameters like Topic Map size, comparison factor between reference and domain corpus, or collocation significance. In the second case, a given Topic Map is enriched by selecting relevant collocations for the topics already in the map, thus enlarging the map.

Figure 4 shows a screenshot of the TopicMapBuilder user interface, which is browser-based and dynamically generated using Java Server



TopicMapBuilder

Version: v1.13 build on 2002/04/05 18:59:16

Auswahl des zu exportierenden Korpus:

Korpus: Auto Motor Sport

Liste: Hitliste

Exportparameter:

Mindestkollokationsstärke: 4 minimale Signifikanz

Faktor gegenüber Allgemeinwortschatz: 4700 50% des Maximalfaktors

Maximale Topic-Anzahl: 400

Häufigkeitsklasse: mindestens: 13 maximal: 6

Export der TopicMap:

Name der TopicMap: TextMining TopicMap

Standard XML generieren: ☑

Figure 4: TopicMapBuilder User Interface

Pages. The knowledge worker producing a Topic Map from the text mining analysis for a given text collection has to fine-tune the currently available system parameters for optimal results in terms of Topic Map size and specificity.

The resulting Topic Map may immediately be exported in the standard XML format for Topic Maps for further usage, e. g. import into in a Topic Map visualization and retrieval tool like the U.S.U. *KnowledgeMiner* [see Gerick 00]. Figure 5 shows a *small* excerpt from a Topic Map generated from a corpus of texts on car technology.

As parameter selection for optimal topic map size (number of topics and associations) can turn out to be a lengthy process, a simple visualization tool is provided to give the user a basic impression of Topic Map quality at first glance. Figure 6 in the appendix shows a sample graph for a car Technology Topic Map, illustrating the relationship between visualization and XML-based Topic Map representation.

The automatically generated topic map is organized as follows: The Topic Map starts with a list of all relevant concepts extracted from the test corpus. These are identified by drawing comparisons with a second, usually much larger reference corpus like the general-purpose corpora discussed in ch. 2.1 (see table 1). The second part of the Topic Map consists of associations identified by collocation analysis as described above. Currently, results still need a great deal of further intellectual refinement (exclusion of irrelevant relations, description of relation types, adding relations not automatically generated), but they can serve as a reliable *basis* for Topic Map construction. As the same baseline process – statistical corpus analysis by using a text mining engine – as described in ch. 2 above is used, the same optimization strategies for filtering collocation sets may be applied.

Conclusion

Text technology methods for Topic Map generation as described in this paper have been successfully applied in industry projects involving domains as diverse as financial and insurance services, chemical and construction engineering, information technology, and general business consulting. Thus, there is a strong indication that the statistical baseline processes are adequate as a general purpose tool in the initial steps of ontology design. While their biggest practical advantage lies in narrowing the gap from ontology design to application, several directions for further research are obvious:

```
<?xml  version="1.0"
       encoding="iso-8859-1"
       standalone="no" ?>
<!DOCTYPE topicmap SYSTEM "map.dtd" >
<topic id="tt.148005"
       categories="tt.206823">
    <comment/>
    <topname>
      <basename>Zahnstangenlenkung
      </basename>
      <dispname>Zahnstangenlenkung
      </dispname>
    </topname>
  </topic>
  <topic id="tt.101202"
         categories="tt.206823">
    <comment/>
    <topname>
      <basename>Querlenkern</basename>
      <dispname>Querlenkern</dispname>
    </topname>
  </topic>

  <!– Association zwischen Querlenkern
      and Zahnstangenlenker -->
<assoc
  id="tt.101202-148005"
  sourcerole="TextMining-association"
  targetrole="TextMining-association"
  sourceid="tt.101202"
  targetid="tt.148005"
  type="ASSOCIATION"
/>
```

Figure 5: Topic Map Excerpt: Coding of Topics and Associations (Corpus on Car Technology)

- Currently, the text mining algorithms are based on different word types in the corpora, accepting synonyms or inflected forms as different concepts. As has been experimentally shown, an ex-ante grouping of surface forms which belong to the same semantic concept is advantageous.
- Likewise, the application of additional linguistic or semantic filters, e. g. leaving out word forms based on their syntactic category or their semantic attribute has a great potential for Topic Map optimization (see [Böhm et al. 02, Heyer, Quasthoff, Wolff 02] for further details).
- A further line of research is the combination of text mining and data mining approaches:

> The fusion of different information sources and analysis methods appears to be a promising path towards better input for ontology design.

Generating Topic Maps for structured information access and retrieval is only one of many possible applications like defining organizational memories (see [Smolnik & Nastansky 02]) or content-oriented structuring of web communities, taking text as representative of typical intentions and goals of community participants.

In more general terms, approaches like the one described here may well give a significant contribution of the *Interspace* as a vision of future distributed information communities (see [Schatz 02]).

Acknowledgements

The authors would like to thank Karsten Böhm of TextTech Ltd., Leipzig, for providing valuable input on the *TopicMapBuilder* tool.

References

[Armstrong 93]. Armstrong, S. (ed.); "Using Large Corpora"; Computational Linguistics 19, 1/2 (1993) [Special Issue on Corpus Processing, repr. MIT Press 1994].

[Biezunski & Newcomb 01] "XML Topic Maps: Finding Aids for the Web"; IEEE Multimedia 8, 2 (2001), 108.

[Böhm et al. 02]. Böhm, K.; Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Topic Map Generation Using Text Mining." In: Journal of Universal Computer Science 8(6) (2002), 623-633.

[Davidson & Harel 96] Davidson, R.; Harel, D; "Drawing Graphs Nicely Using Simulated Annealing."; ACM Transactions on Graphics 15, 4 (1996), 301-331.

[Gerick 00] Gerick, Thomas; "Topic Maps – der neue Standrad für intelligentes Knowledge Retrieval"; Wissensmanagement 2 (2000), 8-12.

[Gruninger & Lee 02]. Gruninger, M.; Lee, J.; Ontology Application and Design". In: Communications of the ACM 45(2) (2002), 39-41.

[Heyer et al. 01] Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, Th.; Wolff, Ch.; "Learning Relations using Collocations"; In: Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, 19-24.

[Heyer, Quasthoff, Wolff 02] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Information Extraction from Text Corpora: Using Filters on Collocation Sets". In: González Rodrígez, M.; Paz Suárez Araujo, C. (edd.) (2002). Proc. LREC 2002. Third Int'l. Conf. On Language Resources and Evaluation, Las Palmas, May 2002, Vol. III, 1103-1107.

[Heyer, Quasthoff, Wolff 00] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Aiding Web Searches by Statistical Classification Tools“. Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz (2000), 163-177.

[Kim 02] Kim, Henry; "Predicting how Ontologies for the Semantic Web will Evolve". In: Communications of the ACM 45(2) (2002), 48-54.

[Krenn 00] Krenn, B.; "Distributional and Linguistic Implications of Collocation Identification." Proc. Collocations Workshop, DGfS Conference, Marburg, March 2000.

[Lemnitzer 98] Lemnitzer, L.; "Komplexe lexikalische Einheiten in Text und Lexikon." In: Heyer, G.; Wolff, Ch. (edd.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 1998, 85-91.

[Maedche & Staab 01] Maedche, A.; Staab, St.; „Ontology Learning for the Semantic Web"; IEEE Intelligent Systems 16, 2 (2001), 72-79.

[Manning & Schütze 99]. Manning, Ch. D.; Schütze, H.; Foundations of Statistical Language Processsing; Cambridge/MA, London: The MIT Press 1999.

[Noy et al. 01] Noy, N. F. ; Sintek, M. ; Decker, St. ; Crubézy, M. ; Fergerson, R. W. ; Musen, M. ; "Creating Semantic Web Contents with Protégé-2000"; IEEE Intelligent Systems 16, 2 (2001), 60-71.

[Quasthoff & Wolff 00] Quasthoff, U.; Wolff, Ch.; "An Infrastructure for Corpus-Based Monolingual Dictionaries." Proc. LREC 2000. Second Int'l. Conf. on Language Resources and Evaluation. Athens, May/June 2000, Vol. I, 241-246.

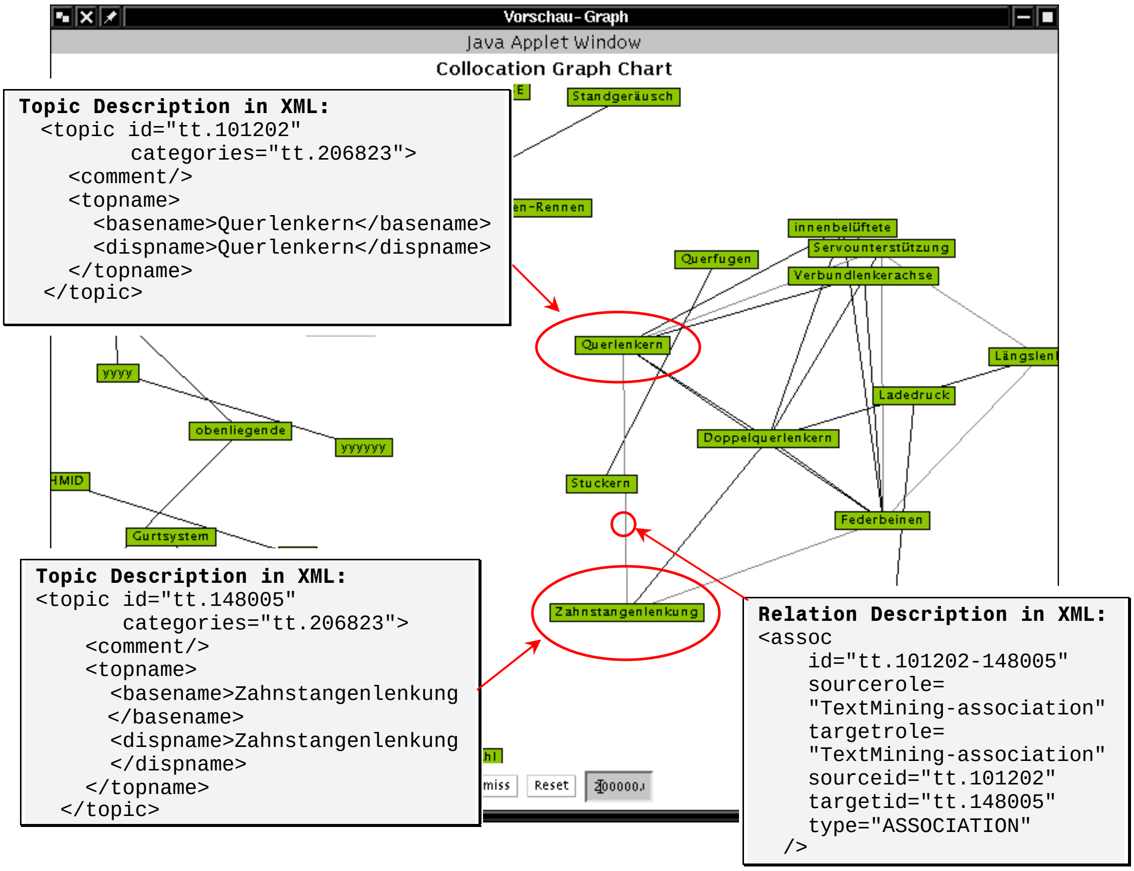
[Schatz 02] Schatz, B.; "The Interspace: Concept Navigation across Distributed Communities"; IEEE Computer 35, 1 (2002), 54-62.

[Smadja 93] Smadja, F.; "Retrieving Collocations from Text: Xtract"; Computational Linguistics 19, 1 (1993), 143-177.

[Smolnik & Nastansky 02] Smolnik, St.; Nastansky, L.; "K-Discovery: Using Topic Maps to Identify Distributed Knowledge Structures in Groupware-based Organizational Memories"; Proc. 35th Annual Hawaiian Int'l Conf. On System Sciences (HICSS-35 '02), Vol. 4.

[Zhou, Booker & Zhang 02] Zhou, L.; Booker, Qu. E.; Zhang, Dongsong; "ROD – Toward Rapid Development for Underdeveloped Ontologies"; Proc. 35th Annual Hawaiian Int'l Conf. On System Sciences (HICSS-35 '02), Vol. 4.

Appendix: Visualization of Raw Topic Maps



Topic Description in XML:

```
  <topic id="tt.101202"
         categories="tt.206823">
    <comment/>
    <topname>
      <basename>Querlenkern</basename>
      <dispname>Querlenkern</dispname>
    </topname>
  </topic>
```

Topic Description in XML:

```
<topic id="tt.148005"
       categories="tt.206823">
   <comment/>
   <topname>
     <basename>Zahnstangenlenkung
     </basename>
     <dispname>Zahnstangenlenkung
     </dispname>
   </topname>
 </topic>
```

Relation Description in XML:

```
<assoc
    id="tt.101202-148005"
    sourcerole=
    "TextMining-association"
    targetrole=
    "TextMining-association"
    sourceid="tt.101202"
    targetid="tt.148005"
    type="ASSOCIATION"
  />
```

Figure 6: Raw TopicMap Visualization