*Journal of Universal Computer Science, vol. 8, no. 6 (2002), 623-633*
*submitted: 4/2/02, accepted: 3/5/02, appeared: 28/6/02 © J.UCS*

# Topic Map Generation Using Text Mining

Karsten Böhm
(TextTech Ltd., Leipzig, Germany
boehm@texttech.de)

Gerhard Heyer
(Leipzig University, Germany
heyer@informatik.uni-leipzig.de)

Uwe Quasthoff
(Leipzig University, Germany
quasthoff@informatik.uni-leipzig.de)

Christian Wolff
(Leipzig University, Germany
wolff@informatik.uni-leipzig.de)

**Abstract:** Starting from text corpus analysis with linguistic and statistical analysis algorithms, an infrastructure for text mining is described which uses collocation analysis as a central tool. This text mining method may be applied to different domains as well as languages. Some examples taken form large reference databases motivate the applicability to knowledge management using declarative standards of information structuring and description. The ISO/IEC Topic Map standard is introduced as a candidate for rich metadata description of information resources and it is shown how text mining can be used for automatic topic map generation.

**Key Words:** Topic Maps, Text Mining, Corpora, Semantic Relations, Knowledge Management
**Categories**: H.3.1, H.3.3, H.5.3, H.3.5, I.2.7, I.7

## 1 Introduction

This paper deals with the automatic extraction of semantic structures from large text collection using the Topic Map ISO standard. The methodology described is based on statistical as well as linguistic analysis routines for text corpora which are described in ch. 1. Chapter 2 gives a short overview of standards for information structuring and introduces Topic Maps and their application to knowledge management. In ch. 3 we discuss our approach for (raw) Topic Map generation and give application examples. Finally advantages and shortcomings of our approach are discussed along with an outlook towards further research.

# 2 Corpus-Based Text Mining

## 2.1 Motivation

Large electronic text collections have become available on the net, fostering research in corpus linguistics and analysis [see Armstrong 93, Manning & Schütze 99]. Furthermore, special collections with specific properties (domain, organisational focus, time) can be set up using internet search agents. At the same time, organisational memories may be constructed automatically from the large amounts of electronically available information in organisations and companies.

## 2.2 An Infrastructure for Text Corpus Analysis

Since 1994 we have been setting up an infrastructure for processing and analysing electronic text corpora [see Quasthoff & Wolff 00]. This infrastructure, which is available on the web (see http://wortschatz.uni-leipzig.de and http://texttech.de) comprises one of the largest online corpora for German, English, and other European languages and offers not only basic information on words and concepts like frequency an basic morphological and grammatical information 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 which is explained ion further detail in the following chapter.

## 2.3 Computation of Significant Collocations

In this paper, the term collocation is used for two or more words with the following statistical property: In a given large corpus, they occur significantly often together within a *predefined window*. Useful windows are

- Next neighbours
- Sentences
- Fixed-size Windows (e. g. *n* word or character distances)
- Documents
- Collections of Documents

We will concentrate on the first two kinds of windows, i. e. next neighbours and sentences, and give only some remarks for very large windows. This selection is motivated by the observation that word neighbourhood as well as sentences boundaries are restrictions that allow for a syntactic as well as semantic interpretation of some kind while fixed size windows impose a restriction that is merely technically motivated.

In calculating collocations, we are interested in the joint occurrence of two given words $A$ and $B$ with probabilities $p_a$ and $p_b$ within a sentence. Let our corpus contain $n$ sentences. For simplicity we will assume that both $A$ and $B$ occur at most *once* in any sentence. This is approximately correct if $A$ and $B$ are not high frequency words.

To measure the surprise of a joint occurrence of A and $B$ we first note that under the assumption of independence of $A$ and $B$ we get a probability of $p_a p_b$ for their joint occurrence in a sample sentence. The number $n$ of sentences in the corpus can be considered as the number of repeated experiments. Using a Poisson distribution [cf. Chung 2000] we get the following approximation for $k$ joint occurrences in the corpus of $n$ sentences, where as usual $\lambda = n\, p_a p_b$:

$$p_k = \frac{1}{k!} \cdot \lambda^k \cdot e^{-\lambda}.$$

As significance measure for collocation we choose the negative logarithm of this probability divided by logarithm of the size of the corpus ($n$):

$$\mathrm{sig(A,B)} = \frac{\lambda - k \cdot \log \lambda + \log k!}{\log n}.$$

The above approximation gives good results for ($k$+1) / $\lambda$ > 10, which is the typical case. For ($k$+1) / $\lambda \leq 10$ we refer to [Läuter & Quasthoff 99]. If, moreover, $k \geq 10$ holds, we might use Stirling's formula to get

$$\mathrm{sig(A,B)} = \frac{k \cdot (\log k - \log \lambda - 1)}{\log n}.$$

This approach was used for calculating the online collocations of German, English, French, Italian, and Dutch corpora of up to 20 million of sentences at *www.wortschatz.uni-leipzig.de.* For an overview of other collocation metrics in the literature see [Lemnitzer 98], [Krenn 00], for an in-depth discussion of the properties of the Poisson approach described above see [Heyer et al. 01] and [Quasthoff & Wolff 02].

### 2.4 Data Overview and Examples

Tables 1 and 2 give an overview of size and contents of our reference corpora and the contents of the German reference corpus, respectively. 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 "Wissen" (German for *knowledge, Fig. 1a)*) and "knowledge" (collocations and visualization, *Fig 1b)*).

| | *German* | *English* | *Dutch* | *French* |
|---|---|---|---|---|
| *word tokens* | 300 m | 250 m | 22 m | 15 m |
| *sentences* | 13.4 m | 13 m | 1.5 m | 860,000 |
| *word types* | 6 m | 1.2 m | 600,000 | 230,000 |

*Table 1: Basic Characteristics of the Corpora*

| Type of data | # of entries |
|---|---|
| Word forms | > 6 Million |
| Sentences | > 25 Millio |
| 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 Millio |

*Table 2: German Corpus Overview*

**Word:** Wissen
**Word Count:** 9443
**Logarithmic Frequency class:** 10 (e. g. $2^{10}$ as seldom as the most frequent word in the corpus)
**Subject Field:** Epistemology Logic
**Morphology:** wiss|en
**Synonyms:** Beschlagenheit, Bildung, Einblick, Einsicht, Erfahrung, Erkenntnis, Faktenwissen, Gelehrsamkeit, Gelehrtheit, Gewißheit […]
**Cross-Reference of:** Bewußtsein, Gewißheit, Kenntnis
**Positive Connotation:** Spezialwissen, Profiwissen, Alleswissen, Sonderwissen, Hauptwissen, Superwissen
**Negative Connotation:** Halbwissen, Pseudowissen
**Base Form:** Wissen
**Inflected Forms:** Wissen, Wissens, Wußten, Wissendes, Wissen
Antonym of: Nichtwissen
-lich-Form of: wissenlich

*Figure 1 a) Basic Statistical Data and Declarative Knowledge for* Wissen

**Top 20 Significant Collocations of *Knowledge*:** Navigator (120), Engineering (78), Systems (78), Base (73), Garden (69), Broker (53), knowledge (50), expert (49), Management (47), Representation (40), System (39), KBMS (35), Expert (32), KnowledgePro (32), expert-system (32), IMKA (31), KEE (31), Garden's (28), Seeker (27), Tool (27), […]

**Significant Left Neighbour Collocations of *Knowledge*:** Managing (36), Applied (34), Discis (20), Carnal (19), Atlantic (17), IntelliCorp's (12), Legal (11), Common (10), Graphic (10), Garden's (9), A&E's (8), Ingres (7), Apollo's (6), MANAGING (6), Contact (5), MEshing (5), Well-Structured (5), pLogic (5), Aided (4), Expert (4), Object-Oriented (4)

**Top 20 Significant Right Neighbour Collocations of *Knowledge*:** Navigator (175), Base (123), Garden (114), Engineering (104), Systems (82), Broker (80), Representation (43), Assets (42), Craft (39), Management (37), Seeker (37), Based (30), Engineers (29), Garden's (28), Pro (28), Access (25), acquisition (24), Network (20), Retrieval (20), Shaper (20), […]

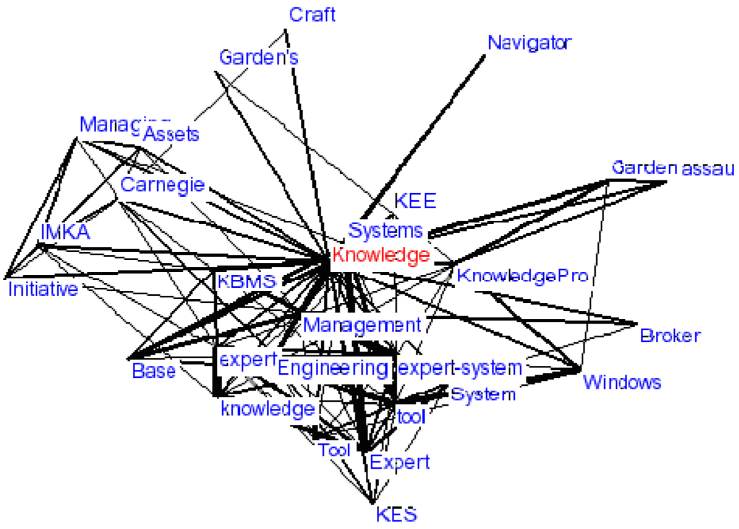

*Figure 1 b) Collocations and Visualization for* Knowledge

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). These data are the starting point for automatic Topic Map generation. 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].

## 2.5 System Architecture

For corpus setup, we employ a flexible software architecture which operates in four basic processing steps:

1. Corpus collection, either by manual provision (e. g. texts on a CD-ROM or collected from intranet servers) or collected by using families of internet search agents.
2. Corpus administration using simple file-system-based tools for handling and converting large amounts of text files.
3. Text analysis including tokenisation and document segmentation, where statistical as well as linguistic algorithms or finding relations like collocations or associations as well as subject categories and hyperonyms and hyponyms are applied.
4. The application layer where the data generated in steps 1-3 are used for different kinds of post-processing (see Heyer, Quasthoff & Wolff 00] for further application details), among them visualisation, indexing, and the generation of knowledge structures using Topic Maps which is described in further detail below.

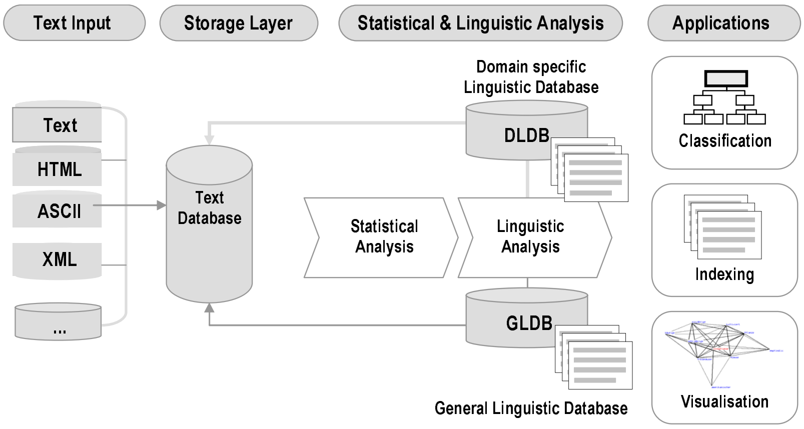

*Figure 2: Overview of System Architecture*

The various tools employed are implemented using standard technologies (C++ and Java as implementation languages, a standard RDBMS as storage layer, XML (eXtensible Markup Language) for information structuring and exchange, JSP and the Tomcat application server for presentation and output). Fig. 2 gives a rough overview of the general system architecture. It should be noted that this basic process of corpus analysis can be applied to any set of text documents and that it is language independent. While we have set up large reference corpora for different languages which represent general knowledge taken from a variety of sources, stressing high-quality periodicals like daily newspapers as well as electronic dictionaries, the same process may also be applied to smaller and more focussed corpora which often represent the organisational knowledge of a certain institution or company. As can be seen in fig. 2, comparison of large reference collections with smaller specific collections is instrumental for selecting key concepts in topic maps.

# 3 Topic Map Generation

## 3.1 Standards for Information Structuring

In recent years, a large number of standards for information and knowledge structuring and description has become available, 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 standardisation efforts:

| *Standard* | *Focus / Application* |
|---|---|
| XML (eXtensible Markup Language) | Document structure (mainly syntactic aspect) |
| Simple HTML Ontology Extension (SHOE) | XML-conformant HTML extension for gathering and representing semantic information about web resources (XML extension) |
| Ontology Markup Language (OML) | Extension of SHOE with different layers, including a RDF mapping |
| RDF (Resource Description Framework / Schema Language) | Description of Metadata, especially for Web Resources; various description schemes or ontologies may be imported or created (RDF Schema Language) |
| Ontology Interchange Language (OIL) | RDF-derived standard for knowledge representation, using a frame-like approach |
| DAML+ OIL (Darpa Agent Markup Language / OIL (Ontology Inference Layer) | Description-logic based standard for ontology engineering, defined on top of RDF |
| Topic Maps (ISO/IEC 13250 standard | Generic standard for document annotation |

*Table 3: XML-based Standards for Information Structuring and Knowledge Representation*

While these different standardisation 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. An in-depth comparison of these standards, unfortunately excluding Topic Maps, is given by [Gómez-Pérez & Corcho 02]

## 3.2 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 occurrences allows for contextualisation of this metadata information. The interrelationship between different topics is formalised by topic association 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 standardised using SGML/XML syntax.

### 3.3 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 analysed prior to topic definition and resource description. This phenomenon is common to both, Topic Maps as well as the Semantic Web initiative and its approaches towards ontology engineering: [Holsapple & Joshi 02: 44] discuss the following methodological approaches towards ontology engineering:

- Inspiration
- Induction
- Deduction
- Synthesis
- Collaboration

These different approaches describe *inherently intellectual processes* which may be simplified or streamlined using appropriate software for ontology editing and optimisation. It might be added, though, that our approach adds a further type of methodology which might tentatively be described as “ontology bootstrapping using text mining” and which can supply and complement the intellectual processes of ontology engineering or Topic Map generation with raw but valuable seed information.

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 modelling tool generating RDF as well as DAML+OIL classes.
- [Maedche & Staab] 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 automatically delivering 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 relation. 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:

- *Topic Map bootstrapping* by corpus comparison, and
- *Topic Map optimisation* for a given Topic Map.

In the first case topic candidates are generating 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 *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.

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<!DOCTYPE topicmap SYSTEM "map.dtd" >
<topicmap name="TextMining TopicMap" id="tt:1">
    <topic id="tt.137523" categories="tt.206823">
    <comment/>
    <topname>
      <basename>Verdichtungsverhaeltnis</basename>
      <dispname>Verdichtungsverhaeltnis</dispname>
    </topname>
  </topic>
    <topic id="tt.72519" categories="tt.206823">
    <comment/>
    <topname>
      <basename>Ladedruck</basename>
      <dispname>Ladedruck</dispname>
    </topname>
  </topic>
<!—Association between Topics 72519 and 137523 -->
  <assoc
    id="tt.72519-137523"
    sourcerole="TextMining-association"
    targetrole="TextMining-association"
    sourceid="tt.72519"
    targetid="tt.137523"
    type="ASSOCIATION"
  />
</topicmap>
```

*Figure 3: Topic Map Excerpt (Car Technology)*

In both cases the resulting Topic Map can immediately be exported in the standard XML format for Topic Maps for further usage, e. g. import into in a Topic Map visualisation and retrieval tool like the U.S.U. *KnowledgeMiner* [see Gerick 00]. Figure 3 shows a small excerpt from a Topic Map generated from a corpus of five volumes of a popular journal on car technology. The results generated by the text mining analysis are formatted according to the specification of the Topic Map DTD (document type definition) and basically consist of two parts:

In the first part, relevant topics are defined as identified after a fine-tuning of extraction parameters like relative frequencies of concepts in a domain specific text, whereas in the second part significant associations between these topics are listed. Using a web-based interface for setting Topic Map extraction parameters along with a simple graph-based visualisation tool for previewing Topic Map results, the knowledge worker may iteratively optimise extraction results prior to further intellectual work on the Topic Map. The results may still need 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.

## 4 Conclusion

The corpus-based methods for Topic Map generation describe in this paper have been successfully applied in industrial settings as diverse as financial and insurance services, chemical engineering or information technology. While their biggest advantage lies in narrowing the gap from Topic Map construction to Topic Map application, several directions for further research are obvious:

- 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 optimisation (see [Heyer et al. 01] for further details). The same holds for a combination of the methods described here with ontology engineering approaches in AI: While the corpus-based approaches are applicable to any domain they may be enhanced be additionally importing existing ontologies.

Generating Topic Maps for structured information access and retrieval is only one of many possible applications like defining organisational memories (see [Smolnik & Nastansky 02]). 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]).

## 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.

[Chung 00]. Chung, Kai Lai "A Course in Probability Theory", Revised Second Edition, New York: Academic Press 2000.

[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.

[Gómez-Pérez & Corcho 02]. Gómez-Pérez, A.; Corcho, O. “Ontology Languages for the Semantic Web”. IEEE Intelligent Systems 17, 1 (2002), 54-60.

[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.

[Holsapple & Joshi 02]. Holsapple, C. W.; Joshi, K. D.; “A Collaborative Approach to Ontology Design.” In: CACM 45, 2 (2002), 42-47.

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

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

[Läuter & Quasthoff 99] Läuter, M.; Quasthoff, U.; “Kollokationen und semantisches Clustering”. In: Proc. 11th. Annual Conference of the Gesellschaft für Linguistische Datenverarbeitung, Frankfurt, July 1999, 34-41.

[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 fort he Semantic Web”; IEEE Intelligent Systems 16, 2 (2001), 72-79.

[Manning & Schütze 99]. Manning, Ch. D.; Schütze, H.; Foundations of Statistical Language Processing; 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 International Conference on Language Resources and Evaluation. Athens, May/June 2000, Vol. I, 241-246.

[Quasthoff & Wolff 02] Quasthoff, U.; Wolff, Ch.; “The Poisson Collocation Measure and its Applications.” In: Proc. International Workshop on Computational Approaches to Collocations, Vienna, July 2002 [to appear].

[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.