E. Milios, Y. Zhang, B. He, L. Dong: ``Automatic Term Extraction and Document Similarity in Special Text Corpora", Proceedings of the 6th Conference of the PacificAssociation for Computational Linguistics (PACLing'03), Halifax,Nova Scotia, Canada, August 22-25, 2003, pages 275-284.

Pacific Association for Computational Linguistics

AUTOMATIC TERM EXTRACTION AND DOCUMENT SIMILARITY IN SPECIAL TEXT CORPORA

E. Milios, Y. Zhang, B. He and L. Dong

Faculty of Computer Science, Dalhousie University, Halifax, Canada B3H 1W5
{*eem,yongzhen,bhe,lid*}*@cs.dal.ca*

This paper confirms that the performance of a state-of-the-art automatic term extraction method on a computer science corpus is similar to previously published performance data on a medical corpus. The extracted terms are then used to estimate the similarity of papers in the computer science corpus using the standard Vector Space Model. The precision of retrieval using a term-based representation is compared with that of a word-based representation, and a link-based similarity metric based on the overlap of the local neighborhoods of the papers in the citation graph. The term-based approach offers comparable performance to the word-based approach, but potentially with a much smaller vocabulary size.

Key words: natural language processing, automatic term extraction, Vector Space Model

1. INTRODUCTION

Automatic term extraction in special text corpora is an interesting problem, which is becoming relevant as literature in specific scientific fields such as medicine, biology and computer science explodes making it difficult to track the evolving terminology in the fields [Kageura and Umino1996]. Early approaches to automatic term extraction were focused on information-theoretic approaches based on mutual information in detecting collocations [Manning and Schuetze1999]. Collocations are expressions that are composed of two or more words, the meaning of which is not easy to guess from the meanings of the component words. There are nuances in the detection of collocation that require linguistic criteria to resolve [Justeson and Katz1995]. Shallow linguistic criteria are based on acceptable sequences of part-of-speech tags. Part-of-speech tagging can be performed automatically [Brill1992]. A key problem is that of nesting, where subsets of consecutive words of terms consisting of multiple words would satisfy the statistical criteria for "termhood", but they would not be called terms.

In the first part of this paper, we describe experiments with a state-of-the-art method, *C-value/NC-value* [Frantzi et al.2000], which combines statistical and linguistic information for automatic term extraction. We applied it to a special text corpus of computer science articles, which is of a different nature from the medical corpus on which the method was originally tested. We confirmed that the performance of the method is equally good on our corpus, and we identified some adjustments that the method required.

In the second part of this paper, we use the terms extracted to estimate the similarity between two documents. We evaluate the quality of the similarity estimation based on terms in an information retrieval context. It is broadly believed that it is difficult to improve upon the bag-of-words representation as far as retrieval performance is concerned by using more sophisticated features or shallow linguistic techniques. Although retrieval based on terms did not show significant improvement over a bag-of-words representation, our long-term objective is to cluster special text corpora into subareas, and automatically generate lexical ontologies from the clusters [Ayad and Kamel2002]. Terms in this context are of interest in themselves, and not purely as a vehicle to information retrieval. We are, furthermore, interested in similarity criteria taking into account proximity of terms [Koubarakis2001], for which again it is essential to work with terms, not words. The use of terms instead of words may also be preferable in information dissemination, where given a database of profiles (of

users) and a document, the profiles that match the document must be identified efficiently [Koubarakis et al.2002].

The contribution of this work has been to confirm that a state-of-the-art method for automatic term extraction performs well in different special text corpora, and that similarity estimation based on terms performs at least as well as the standard bag-of-words representation in a document retrieval context. We further compared the performance of term-based retrieval with that of a method based only on links in the citation graph constructed based on the references (and citations) of the computer science articles considered [Lu et al.2001]. We observed overlap in the results from the term-based and the link-based methods, but also relevant articles returned by one but not the other methods. So it appears that the methods need to be combined in order to get the best retrieval performance.

Section 2 describes the *C-value/NC-value* method for automatic term extraction and its evaluation. Section 3 describes the similarity estimation based on terms, including the vocabulary and parameter selection, and the experimental results from the term-based and word-based methods, and compares them with those of the link-based method applied to the same corpus. Section 4 discusses the approach and directions for future research.

2. AUTOMATIC TERM EXTRACTION

This section describes how *C-value/NC-value* method works on a computer science corpus to extract multi-word terms automatically. We experimentally confirmed that the performance of the method on the computer science corpus is as good as that on a medical corpus.

2.1. C-value Method

C-value [Frantzi et al.2000] is a domain-specific method used to automatically extract multi-word terms. It aims to get more accurate terms than those obtained by the pure frequency of occurrence method, especially terms that may appear as nested within longer terms, for example "MUTUAL INFORMATION" nested in "MUTUAL INFORMATION FEATURE EXTRACTION CRITERIA". The *C-value* method combines linguistic knowledge (which consists of part-of-speech tagging, linguistic filters, stop list, etc.) and statistical information, to obtain a termhood measure called *C-value*. The larger this value, the higher the probability for a candidate term to be a real term.

A linguistic filter is used to extract word sequences likely to be terms, such as noun phrases and adjective phrases. In this project we used three kinds of filters as follows:

- $\text{Noun}^{+}\text{Noun}$
- $(\text{Adj}|\text{Noun})^{+}\text{Noun}$
- $((\text{Adj}|\text{Noun})^{+}|((\text{Adj}|\text{Noun})^{*}(\text{NounPrep})^{?})(\text{Adj}|\text{Noun})^{*})\text{Noun}$

Obviously, different filters have different results on precision and recall. The first one is a "closed filter" [Frantzi et al.2000], i.e. the candidate terms extracted using it will have higher precision and lower recall. The last two are "open filters" [Frantzi et al.2000] which have the tendency to get higher recall but lower precision.

C-value method shows improved performance in non-nested terms because it takes a term's length into consideration. The *C-value* method is applied to a computer science corpus and candidate terms are extracted and ordered by the *C-value*.

2.2. NC-value Method

In this subsection we discuss how to incorporate context information into Automatic Term Extraction.

Term context words. We often identify the meaning of a phrase by using its context information. In Automatic Term Extraction, information from a set of "modifiers" (mainly nouns, verbs and adjectives) is used to distinguish between terms and non-terms [Frantzi et al.2000]. We create a list of "important" *term context words* from a set of candidate terms extracted by the *C-value* method. Term context words are those that appear in the vicinity of candidate terms, i.e. nouns, verbs and adjectives that either precede or follow the candidate term.

Research in [Frantzi et al.2000] shows that term context words are domain (corpus) dependent and assigning high values to words that tend to appear with terms will improve the extraction of terms.

NC-value. The *NC-value* is an extension to *C-value*, which incorporates context information into term extraction. The method involves three stages:

1. Domain experts manually label all real terms in the top 10% candidate terms ranked by *C-value* method for the extraction of term context words.
2. Context words are extracted from each term obtained in the previous stage and record their appearance times. Then each word is assigned a weight.
3. The candidate term list extracted by *C-value* is re-ranked with the incorporation of context information acquired from the second stage.

2.3. C-value/NC-value Evaluation

The medical corpus consists of 810,719 words of a few thousand short eye-pathology medical records in the corresponding application [Frantzi et al.2000]. In order to conduct the same experiments as done in [Frantzi et al.2000] and make reasonable comparisons of performance on both corpora, we selected 1000 papers from the neural network corpus (a total of 10426 papers) to construct a mini corpus. From each of these 1000 papers, the first 820 words were extracted, leading to a total of 820,000 words in the mini corpus. Because the selection of articles under the neural network topic was automated, and terminology usage in computer science is probably less precise than in eye-pathology diagnoses, we expect the computer science corpus to be less coherent than the medical corpus, with more terms and fewer instances of each term present in the corpus.

Since Automatic Term Extraction techniques are empirical [Kageura and Umino1996], we evaluated the performance of *C-value/NC-value* method on the computer science corpus in terms of *precision* and *recall*.

Table 1 shows the precision results of *C-value* method obtained on the computer science corpus and the medical corpus.

The precisions from both corpora give very similar conclusion:

- Filter one gains higher precision than filter two and filter three.
- *C-value* achieves similar precision on the computer science corpus as on the medical corpus. A formal examination of the small difference can be conducted using a standard statistical test in the future research.

Recall is the ratio of the number of real terms extracted to the number of real terms in

corpus	1^{st} **filter**	2^{nd} **filter**	3^{rd} **filter**
computer science	35.73%	32.95%	30.37%
medical	38%	36%	31%

Table 1. Precision of *C-value* vs. frequency on the computer science corpus and the medical corpus. Precision is the ratio of the number of real terms extracted to the number of candidate terms extracted. Three linguistic filters were used in the test.

the corpus. Given the fact that it is very time-consuming for domain experts to look through the corpus and obtain all the real terms, we instead calculate recall in three steps. First we use a frequency threshold of 3 to get all the candidate terms. Second, we count the number of real terms in this candidate term list. Third, we use an additional *C-value* threshold of 0 to filter the term list obtained from step 2. Now we can calculate recall as the ratio of the number of terms after step2 to the number of terms after step 1 [Frantzi et al.2000].

Table 2 shows the recall numbers of *C-value* on both corpora.

corpus	1^{st} **filter**	2^{nd} **filter**	3^{rd} **filter**
computer science	97.57%	96.88%	94.42%
medical	98.22%	97.41%	97.47%

Table 2. Recall of *C-value* on the computer science corpus and the medical corpus.

As we can see, recall of *C-value* on the computer science is a little lower than that on the medical corpus, but the difference is very small. Moreover, recall of *C-value* does not lose much using the first two linguistic filters for both corpora.

The overall precision and recall of *NC-value* are the same as that of *C-value* because *NC-value* just re-ranks the candidate term list extracted by *C-value* without adding or deleting any candidate terms [Frantzi et al.2000]. However, we want to see the precision of *NC-value* on different intervals of candidate term list, and compare with that of *C-value* and frequency of occurrence.

Figure 1 and Figure 2 show the precision obtained from the computer science corpus and the medical corpus, respectively. Both of them use linguistic filter 3, the most open filter. Both horizontal axes are divided into several intervals to show their distributed precision. The breakdown of the horizontal axis into intervals is corpus dependent and is based on the requirement that each interval contains about the same number of candidate terms [Frantzi et al.2000]. The intervals for the medical corpus used in [Frantzi et al.2000] are [top-40], (40-10], (10-4), [4-1], while the intervals for the computer science corpus are [top-30], (30-8], (8-4), [4-3]. The vertical axes show the precision on the terms belonging to the corresponding interval.

As shown in Figure 1 and Figure 2, the precision result from the computer science corpus confirms that of the medical corpus: *NC-value* achieves higher precision on the candidate terms at the top of the list than the other two methods. It increases the concentration of real terms at the top of the list. More precisely, it brings a 5% increase of precision for the first two intervals. For the last two intervals, we observe a drop in precision due to the increase of precision for the first two intervals.

The differences in precision values between the computer science corpus and the medical

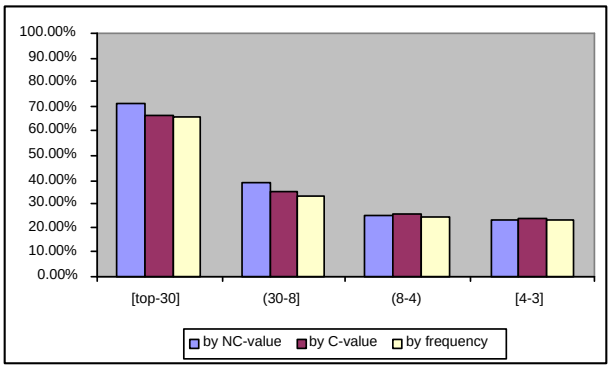

Figure 1. Distributed precision on the computer science corpus

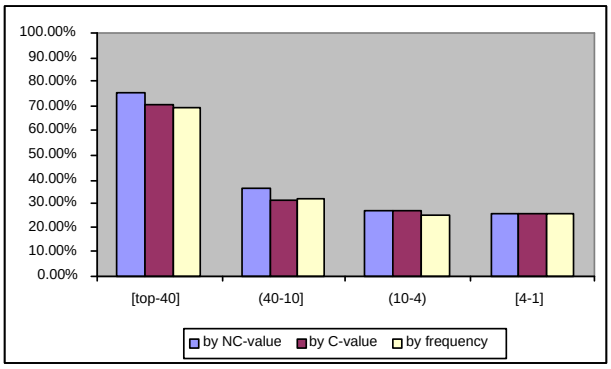

Figure 2. Distributed precision on the medical corpus

corpus may be explained by the medical corpus being likely to make more focused use of fewer terms.

Our experiments showed that *C-value/NC-value* performs on the computer science corpus as well as on the medical corpus.

3. DOCUMENT SIMILARITY ASSESSMENT

In this section, the terms extracted by the method presented in the previous section will be used to form a *document vector space* for defining a similarity measure and the precision of a retrieval task based on this measure will be analyzed.

The Vector Space Model is widely used for the measurement of similarity between documents [Manning and Schuetze1999] because of its conceptual and computational simplicity. Documents and queries are represented as vectors in a vector space, where the dimensions correspond to "features" (words or terms).

In this research, we applied the following equation [Manning and Schuetze1999] to define the term weight.

$$weight(i,j) = \begin{cases} (1 + \log tf_{i,j}) \cdot \log \frac{N}{df_i} & if \quad tf_{i,j} \geq 1 \\ 0 & if \quad tf_{i,j} = 0 \end{cases} \tag{1}$$

where
$tf_{i,j}$ is the frequency of term i in document j,
df_i is the number of documents in which term i occurs, and
N is the total number of documents in the corpus.

Documents are ranked in the vector space model by measuring their similarities with the query vector using the cosine distance.

$$cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \sqrt{\sum_{i=1}^{n} d_i^2}} \tag{2}$$

where $\vec{q}$ and $\vec{d}$ are n-dimensional vectors.

Our objective is to evaluate the use of terms as features in a vector space model.

In the similarity assessment, first we choose 9 query papers used in [Lu et al.2001] for document similarity assessment with link-based method. These papers were selected based on several criteria for experimental purpose to obtain representative results. For example, we may choose papers published in different times or with different significance.

Next we estimated the 10 most similar papers for each of the given 9 query papers using terms extracted from the corpus. Each of the top 10 similar papers was evaluated as related (score 1), somewhat related (score 0.5) or not related (score 0) to the query paper.

3.1. Evaluation Methods

Three different methods were used to evaluate document similarity: *term-based*, *word-based*, and *link-based*.

In the term-based method, we generated the corpus terms from full papers sorted by frequency of occurrence, *C-value* and *NC-value*. There were 189,043 candidate terms extracted from the whole corpus with a specific linguistic filter, but not all the terms are suitable for information retrieval [van Rijsbergen1999]. We specified two cut-offs to exclude the most frequent and the least frequent terms. Consequently, we extracted a subset of 6100 terms with frequency between them [van Rijsbergen1999]. We experimented with both frequency/ *C-value/NC-value* and document frequency to choose the cut-offs, and with different cut-off points. *Term document frequency* is the number of documents in which the term occurs [Manning and Schuetze1999]. Cutoffs were determined empirically, as is common practice in information retrieval.

In the word-based method, we extracted all the nouns from the corpus as features. The number of these nouns is 11060, almost two times of the term size.

The top 10 similar papers for each of 9 query papers were judged by two domain experts, and were assigned a score 1 (related), 0.5 (somewhat related) or 0 (not related).

3.2. Precision comparison among methods

As shown in Figure 3, the term-based method gave on the average somewhat better precision than the word-based and link-based methods.

To determine the statistical significance of the above observations, we performed a two-factor Analysis of Variance with replications on the raw scores from the above experiments.

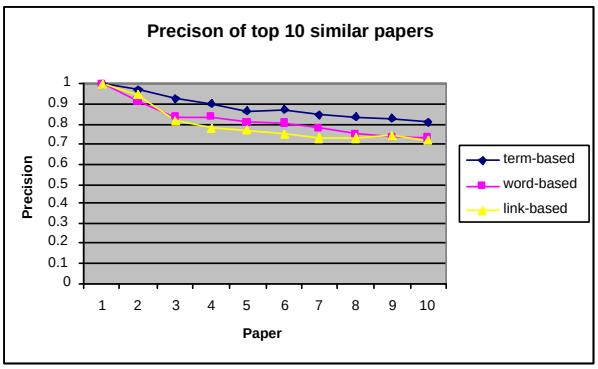

FIGURE 3. Precision comparison of three methods: term-based, word-based, and link-based. The precision for paper ranked N is the fraction of papers deemed related to the query paper by the experts among the top N most similar papers returned by the algorithm, averaged over the 9 query papers.

We show the F-statistic and P-value for each pairwise test in Table 3. As we can see, the significance level of the difference between the term-based and word-based methods is 87% (i.e. there is probability 13% that the observed difference came about by chance). The significance level of the difference between the term-based and link-based methods is 92% (i.e. there is probability 8% that the observed difference came about by chance). There is no significant difference between the word-based and the link-based methods.

We would run evaluation experiments with smaller vocabularies to find the limit where the word method breaks down because of an impoverished vocabulary, and expect that for equal size vocabularies, term clearly beats word.

	term-based	word-based
word-based	$F_{1,162} = 2.28$ $Pvalue = 0.13$	
link-based	$F_{1,162} = 3.10$ $Pvalue = 0.08$	$F_{1,162} = 0.04$ $Pvalue = 0.85$

TABLE 3. Pairwise ANOVA results for the three experiments.

3.3. Complementarity of methods

Term-based vs. word-based. The word-based and term-based methods complement each other by producing different sets of related papers, see Figure 4. Averaged over 9 query papers, they had 4.4 relevant papers in common against the top 10 similar papers and for the remaining non-common papers, 3.8 papers were judged as relevant with term-based method and 3 papers were judged as relevant with word-based method.

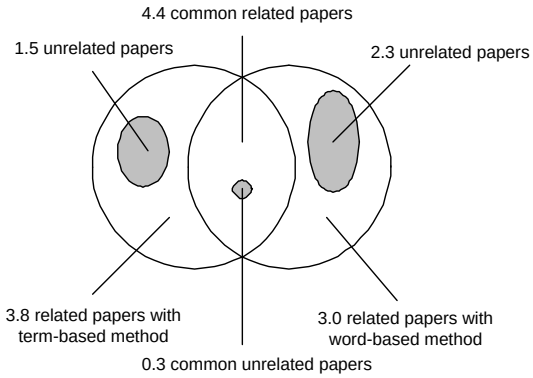

Figure 4. Venn diagram for the complementarity of the results from the term-based and word-based methods

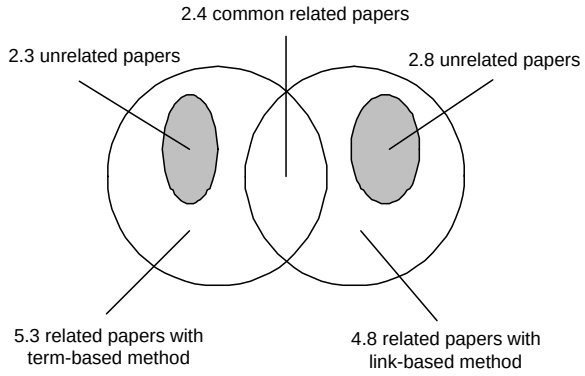

Figure 5. Venn diagram for the complementarity of the results from the term-based and link-based methods

Term-based vs. link-based. The link-based and term-based methods complement each other by producing different sets of related papers, see Figure 5. Averaged over 9 query papers, they had 2.4 relevant papers in common against the top 10 similar papers and for the remaining non-common papers, 5.3 papers were judged as relevant with term-based method and 4.8 papers were judged as relevant with link-based method.

The term-based method can get higher precision but needs time to preprocess the texts and build an inverted index. Term- and link-based methods can be used together to gain higher precision and attract more similar papers to the top similar paper list.

Word-based vs. link-based. Figure 6 demonstrates that the link-based and word-based methods complement each other too by producing different sets of related papers. Averaged over 9 query papers, they had 2.8 relevant papers in the 2.9 common papers against the top 10 similar papers and for the remaining non-common papers, 4.6 papers were judged as relevant with word-based method and 4.4 papers were judged as relevant with link-based

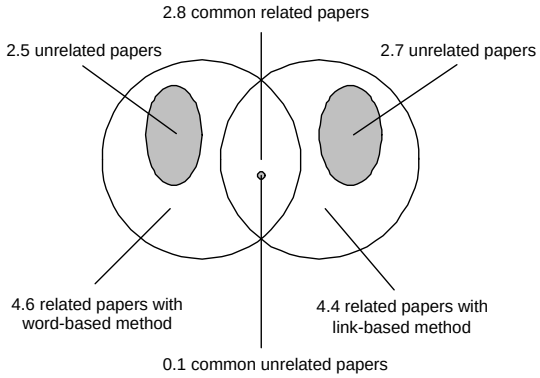

FIGURE 6. Venn diagram for the complementarity of the results from the word-based and link-based methods

method.

In future research, it would be useful to compare the ranked list of more than than the top 10 most related papers (according to each similarity measure), to see whether the above pattern of complementarity of the returned papers persists.

4. CONCLUSION

In this paper, we confirmed that the performance of the *C-value/NC-value* method in automatic term extraction on a corpus consisting of computer science articles is as good as that on the medical corpus, on which the method was tested by its original authors.

Furthermore, we used automatically extracted terms as features in an information retrieval task, in which we were given a corpus of papers and we were searching for the most similar papers in the corpus to a given query paper. Terms of intermediate frequency were selected, according to standard practice in information retrieval. We compared the precision obtained using terms, words and a link-based method that is based entirely on the information encoded in the citation graph. Precision of term-based retrieval appeared to be higher than word-based and link-based retrieval. The papers returned by the methods have substantial overlap, but there are several papers returned by one method and not the others. So from an information retrieval perspective, the methods are complementary and using them together would be beneficial.

Future research involves several directions.

- Efficient implementation using the proper data and file structures for handling large special text corpora and number of terms.
- Integration of term-based and link-based methods for document retrieval.
- Demonstration of how term extraction varies with topic domain and text genre, possibly leading to term extraction techniques specialized for various types of texts.
- Applying term-based similarity to clustering of special text corpora, aiming to automatically discover hierarchical organizations of scientific disciplines and the associated induced lexical ontologies [Ayad and Kamel2002, Clifton et al.1999, Maynard and Ananiadou2000a, Maynard and Ananiadou2000b], and knowledge mining in the Web [Mima et al.1999].

REFERENCES

[Ayad and Kamel2002] H. Ayad and M. Kamel. 2002. Topic discovery from text using aggregation of different clustering methods. In *AI'2002: The Fifteenth Canadian Conference on Artificial Intelligence*, 27-29 May, 2002.

[Brill1992] E. Brill. 1992. A simple rule-based part of speech tagger. In *3rd Conference on Applied Natural Language Processing (ANLP-92)*, pages 152–155.

[Clifton et al.1999] C. Clifton, R. Cooley, J. Zytkow, and J. Rauch. 1999. Topcat: data mining for topic identification in a text corpus. In *Principles of Data Mining and Knowledge Discovery. Third European Conference, PKDD'99*, pages 174–183.

[Frantzi et al.2000] K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2):117–132.

[Justeson and Katz1995] J. Justeson and S. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *In Natural Language Engineering*, 1:9–27.

[Kageura and Umino1996] K. Kageura and B. Umino. 1996. Methods of automatic term recognition -a review-. *Terminology*, 3(2):259–289.

[Koubarakis et al.2002] M. Koubarakis, T. Koutris, P. Raftopoulou, and C. Tryfonopoulos. 2002. Efficient dissemination of textual information using the boolean model. In *2nd Hellenic Conference on Artificial Intelligence*.

[Koubarakis2001] M. Koubarakis. 2001. Boolean queries with proximity operators for information dissemination. In *International Workshop on FOUNDATIONS OF MODELS FOR INFORMATION INTEGRATION (FMII-2001), as the 10th Workshop in the Series Foundations of Models and Languages for Data and Objects (FMLDO), immediately after VLDB-2001*.

[Lu et al.2001] W. Lu, J. Janssen, E. Milios, and N. Japkowicz. 2001. Node similarity in networked information spaces. *Technical Report CS-2001-03 Dalhousie University*, September.

[Manning and Schuetze1999] C. Manning and H. Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, Massachusetts.

[Maynard and Ananiadou2000a] D. Maynard and S. Ananiadou. 2000a. Creating and using domain-specific ontologies for terminological applications. *Proceedings of Second International Conference on Language Resources and Evaluation, Athens*.

[Maynard and Ananiadou2000b] D. Maynard and S. Ananiadou. 2000b. Identifying terms by their family and friends. *The 18th International Conference on Computational Linguistics, COLING 2000*.

[Mima et al.1999] H. Mima, S. Ananiadou, and J. Tsujii. 1999. A web-based integrated knowledge mining aid system using term-oriented nlp. *Proceedings of Natural Language Processing Pacific Rim Symposium 99, Beijing*.

[van Rijsbergen1999] C. van Rijsbergen. 1999. *Information Retrieval*. `http://www.dcs.gla.ac.uk/~iain/keith/index.htm`, 2nd ed., last accessed on April 17, 2002.