International Journal of Medical Informatics (2006) **75**, 488—495

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library

Imad Tbahriti[a,c], Christine Chichester[b], Frédérique Lisacek[b,c], Patrick Ruch[a,*]

[a] *SIM, University and Hospitals of Geneva, 24 Micheli du Crest, 1211 Geneva, Switzerland*
[b] *Geneva Bioinformatics (GeneBio) SA, 25 Avenue de Champel, Geneva, Switzerland*
[c] *Swiss Institute of Bioinformatics, Geneva, Switzerland*

Received 26 April 2005; received in revised form 24 June 2005; accepted 28 June 2005

KEYWORDS
Argumentation;
MEDLINE;
Citation;
Information storage and retrieval;
Related article search

Summary The aim of this study is to investigate the relationships between citations and the scientific argumentation found abstracts. We design a related article search task and observe how the argumentation can affect the search results. We extracted citation lists from a set of 3200 full-text papers originating from a narrow domain. In parallel, we recovered the corresponding MEDLINE records for analysis of the argumentative moves. Our argumentative model is founded on four classes: PURPOSE, METHODS, RESULTS and CONCLUSION. A Bayesian classifier trained on explicitly structured MEDLINE abstracts generates these argumentative categories. The categories are used to generate four different argumentative indexes. A fifth index contains the complete abstract, together with the title and the list of Medical Subject Headings (MeSH) terms. To appraise the relationship of the moves to the citations, the citation lists were used as the criteria for determining relatedness of articles, establishing a benchmark; it means that two articles are considered as ''related'' if they share a significant set of co-citations. Our results show that the average precision of queries with the PURPOSE and CONCLUSION features is the highest, while the precision of the RESULTS and METHODS features was relatively low. A linear weighting combination of the moves is proposed, which significantly improves retrieval of related articles.

1. Introduction

Numerous techniques help researchers locate relevant documents in an ever-growing mountain of scientific publications. Among these techniques is the analysis of bibliographic information, which

[*] Corresponding author. Tel.: +41 22 372 61 64.
E-mail address: patrick.ruch@sim.hcuge.ch (P. Ruch).

1386-5056/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2005.06.007

can identify conceptual connections between large numbers of articles. Although helpful, most of these systems deliver masses of documents to the researcher for analysis, which contain various degrees of similarity. This paper introduces a method to determine the similarity of a bibliographic co-citation list, that is the list of citations that are shared between articles, and the argumentative moves of an abstract in an effort to define novel similarity searches.

Authors of biological papers develop arguments and present the justification for their experiments based on previously documented results. These results are represented as citations to earlier scientific literature and establish the links between old and new findings. The assumption is that the majority of scientific papers employing the same citations depict related viewpoints. The method described here is applied to improve retrieval of similar articles based on co-citations, but other applications are possible, such as information extraction and term normalization as explored in Nakov et al. [1]. Documents that should be conceptually correlated due to bibliographic relatedness but which propose different or novel arguments are often not easily located in the majority of bibliographically correlated articles. Our system can be tuned to identify these documents. Conversely, such a system could also be used as a platform to aid authors by means of automatic assembly or refinement of their bibliographies through the suggestion of citations coming from documents containing similar arguments.

The rest of this paper is structured as follows: Section 2 describes the background related to experiments using citations or argumentation that compare aspects connected to the logical content of publications. Section 3 details the method and the generation of the different indexes used in our analyses, e.g. the citation index, the four argumentative indexes and the abstract index (abstract, title and keywords). Section 4 presents the results of the evaluations we performed. Section 5 closes with a summary of the contribution of this work, limitations and future work.

2. Background

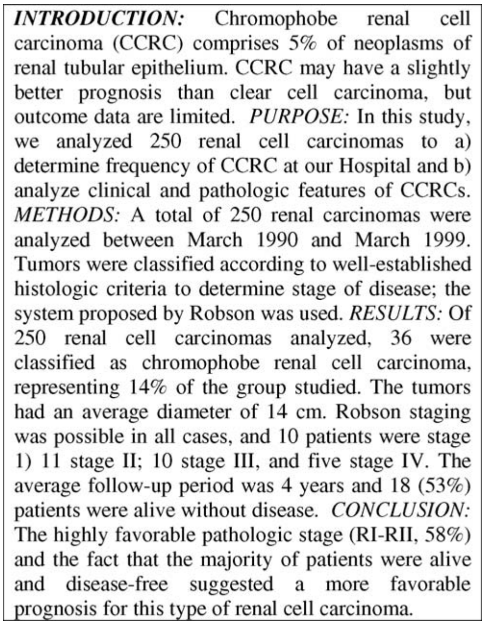
Digital libraries aim at structuring their records to facilitate user navigation. Interfaces visualizing overlapping relationships of the standard library fields such as author and title in document collections are usually the most accessible to the user. Beyond these well-known targets, researchers (see Ref. [2], or [3], for a survey) interested in information extraction and retrieval for biomedical applications have mostly focused on studying specific biological interactions [4—7] and related entities [8—11,7] or using terms in biomedical vocabularies [12—15]. The use of bibliographical and argumentative information [16] has been less well studied by researchers interested in applying natural language processing to biomedical texts.

2.1. Citations

Originating from bibliometrics, citation analysis [17] has been used to visualize a field via a representative slice of its literature. Co-citation techniques make it possible to cluster documents by scientific paradigm or hypothesis [18]. Braam et al. [19] have investigated co-citation as a tool to map subject-matter specialties. They found that the combination of keyword analysis and co-citation analysis was useful in revealing the cognitive content of publications. Peters et al. [20] further explored the citation relationships and the cognitive resemblance in scientific articles. Word profile similarities of publications that were bibliographically coupled by a single, highly cited article were compared with publications that were not bibliographically coupled to that specific article. A statistically significant relationship has been established between the content of articles and their shared citations. This result will serve as basis to establish our benchmark without relevance judgments [21,22]. Thus, we define a new concept of relevance, which is not based on the judge's subjectivity but on bibliographical contents. A retrieved article is related to another article — used as query — if these two articles share a significant number of citations.

2.2. Argumentation in biomedical abstracts

Scientific research is often described as a problem solving activity. In full text scientific articles this problem—solution structure has been crystallized in a fixed presentation known as INTRODUCTION, METHODS, RESULTS and CONCLUSION. This structure is often presented in a much-compacted version in the abstract and it has been clearly demonstrated by Ehrler et al. [15] that abstracts contain a higher information density than full text. Correspondingly, the four-move problem-solving structure (standardized according to ISO/ANSI guidelines) has been found quite stable in scientific reports [24]. Although the argumentative structure of an article is not always explicitly labeled, or can be labeled using slightly different



INTRODUCTION: Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited. *PURPOSE:* In this study, we analyzed 250 renal cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs. *METHODS:* A total of 250 renal carcinomas were analyzed between March 1990 and March 1999. Tumors were classified according to well-established histologic criteria to determine stage of disease; the system proposed by Robson was used. *RESULTS:* Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied. The tumors had an average diameter of 14 cm. Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV. The average follow-up period was 4 years and 18 (53%) patients were alive without disease. *CONCLUSION:* The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.

Fig. 1 Example of an explicitly structured abstract in MEDLINE. The four-class argumentation model is sometimes split into classes that may carry slightly different names, as illustrated in this example by the INTRODUCTION marker.

markers (as seen in Fig. 1), a similar implicit structure is common in most biomedical abstracts [25]. Therefore, to find the most relevant argumentative status that describes the content of the article, we employed a classification method to separate the content dense sentences of the abstracts into the argumentative moves.

3. Methods

We established a benchmark based on citation analysis to evaluate the impact of using argumentation to find related articles. In information retrieval, benchmarks are developed from three resources: a document collection, a query collection and a set of relevance rankings that relates each query to the set of documents. Existing information retrieval collections normally contain user queries composed of only a few words [26]. These short queries are not suitable for evaluating a system tailored to retrieve articles with similar citations. Therefore, we have created the collection and tuned the system to accept long queries such as abstracts (Fig. 2).

3.1. Data acquisition and citation indexing

All the data used in these experiments were acquired from MEDLINE using the PubMed interface.

3.1.1. Document collection

The document set was obtained from PubMed by executing a set of Boolean queries to recover articles related to small active peptides from many animal species excluding humans. These peptides hold the promise of becoming novel therapeutics. The set consisted of 12500 documents, which were comprised of abstract, title and MeSH terms. For 3200 of these documents we were able to recover the full text including the references for citation extraction and analysis.

3.1.2. Queries

Following statistical analysis confirmed by Buckley and Voorhees [27], four sets of 25 articles were selected from the 3200 full text articles. The title, abstract and MeSH terms fields were used to construct the queries. For testing the influence the argumentative move, the specific sentences were extracted and tested either alone or in combination with the queries that contained the title, abstract and MeSH terms.

3.1.3. Citation analysis

Citation lists were automatically extracted from 3200 full-text articles that were correspondingly represented within the document set. This automatic parsing of citations was manually validated. Each citation was represented as a unique ID for comparison purposes. Citation analysis of the entire collection demonstrated that the full-text articles possessed a mean citation count of 28.30 ± 24.15 (mean ± S.D.) with a 95% CI = 27.47—29.13. Within these records the mean co-citation count was 7.79 ± 6.99 (mean ± S.D.) with a 95% CI = 7.55—8.03. As would be expected in a document set which contains a variety of document types (reviews, journal articles, editorials), the standard deviations of these values are quite large.

3.1.4. Citation benchmark

For each set of queries, a benchmark was generated from the 10 cited articles that contained the greatest number of co-citations in common with the query. For the benchmark, the average number of cited articles that have more than nine co-citations was 15.70 ± 6.58 (mean ± S.D.). Query sets were checked to confirm that at least one sentence in each abstract was classified per argumentative class. We observe that the standard deviation in the query set is inferior to the standard deviation in the complete corpus. This is mainly due to the fact that the co-citation threshold tends to exclude articles that have very few citations, such as editorials. Thus, the average number (about 16)

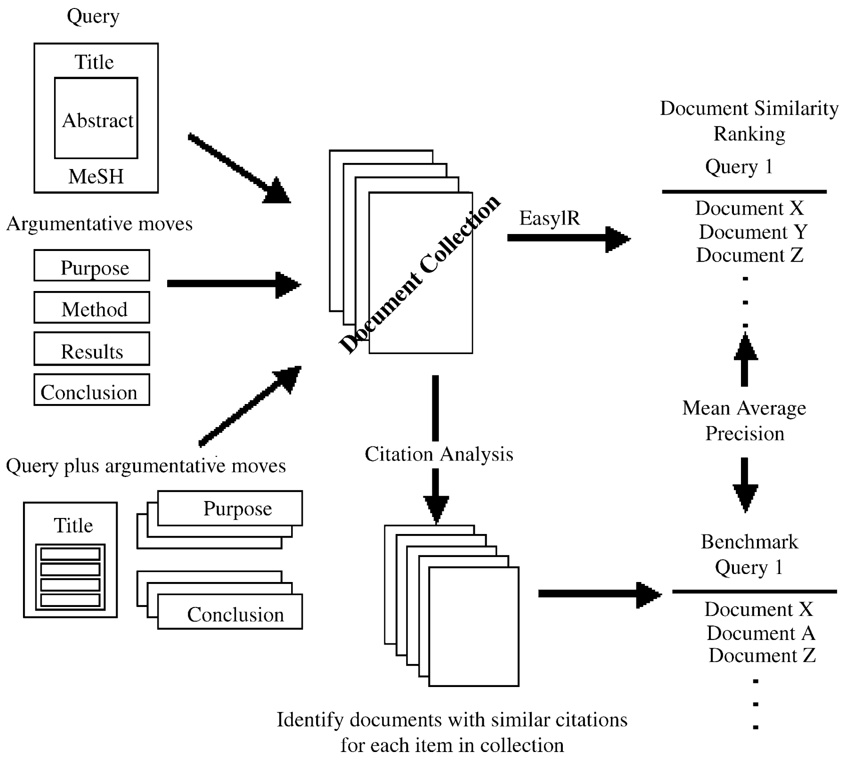


Fig. 2 Flowchart for the chain of experimental procedures. The benchmark was assembled from citations shared between documents and compared to the document similarity ranking of EasyIR.

is representative of regular articles, such as those of interest when searching for similar articles.

3.2. Metrics

The main measure for assessing information retrieval engines is mean average precision (MAP). MAP is the standard metric, although evidences have been provided that it tends to hide minor differences in ranking [28]; therefore complementary metrics, such as the precision at 5 or 10 retrieved articles are often provided as well.

3.3. Text indexing

For indexing, we used the easyIR system,[1] which implements standard vector space IR schemes as well as more advanced models such as the l(n)L2 weighting [29], which are already effective without pseudo-relevance feedback [30].

However, we restrict our investigation to the standard vector space model in order to limit the complexitiy of the experiments. The term-weighting schema composed of combinations of term frequency, inverse document frequency and length normalization was varied to determine the most relevant output ranking. Table 1 gives the most common term weighting factors (atc.atn, ltc.atn); the first letter triplet applies to the document, the second letter triplet applies to the query [31].

3.4. Argumentative classification

The classifier segmented the abstracts into four argumentative moves: PURPOSE, METHODS,

Table 1 Weighting parameters as in SMART

First letter	$f(tf)$
Term frequency	
n (natural)	tf
1 (logarithmic)	$1+\log(tf)$
a (augmented)	$0.5+0.5\times(tf/\max(tf))$
Second letter	$f(1/df)$
Inverse document frequency	
n (no)	1
t (full)	$\log(N/df)$
Third letter	$f(\text{length})$
Normalization	
n (no)	1
c (cosine)	$\sqrt{\rho_1^2+\rho_2^2+\cdots+\rho_n^2}$

[1] http://lithwww.epfl.ch/~ruch/softs/softs.html.

Table 2 Confusion matrices for each argumentative class: results are in percent

	PURPOSE	METHODS	RESULTS	CONCLUSION
PURPOSE	93.55	0	3.23	3
METHODS	8	81	8	3
RESULTS	7.43	5.31	74.25	13.01
CONCLUSION	2.27	0	2.27	95.45

RESULTS and CONCLUSION. The classification unit is the sentence which means that abstracts are preprocessed using an ad hoc sentence splitter. The confusion matrix for the four argumentative moves generated by the classifier is given in Table 2. This evaluation used explicitly structured abstracts; therefore, the argumentative markers were removed prior to the evaluation. Fig. 3 shows the output from the classifier, when applied to the abstract shown in Fig. 1.

After extraction, each of the four types of argumentative moves was then used for indexing, retrieval and comparison tasks.

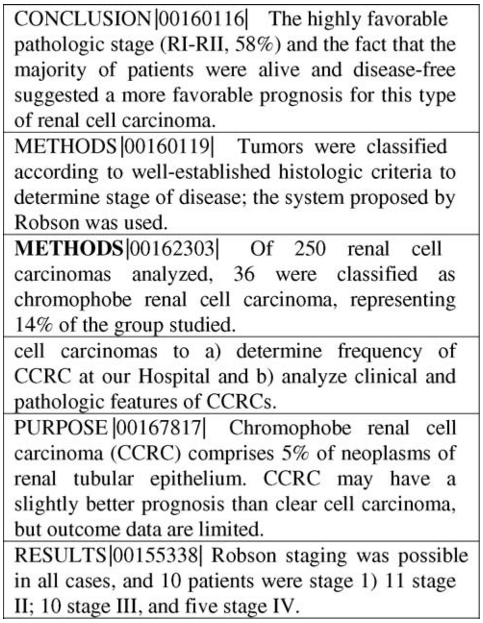
3.5. Argumentative combination

We adjusted the weight of the four argumentative moves, based on their location and then combined them to improve retrieval effectiveness. The query weights were recomputed as indicated in Eq. (1).

$$W_{\mathrm{new}} = W_{\mathrm{old}} S_{\mathrm{c}} k_{\mathrm{c}} \qquad (1)$$

where $c \in \{$PURPOSE; METHODS; RESULTS; CONCLUSION$\}$; W_{old} is the feature weight as given by the query weighting (ltc); S the *normalized* score attributed by the argumentative classifier to each sentence in the abstract. This score is attributed to each feature appearing in the considered segment. k is a constant for each value of c. The value is set empirically using the tuning set (TS). The initial value of k for each category is given by the distribution observed in Table 4 (i.e., 0.625, 0.164, 0.176, 0.560 for the classes, PURPOSE, METHODS, RESULTS and CONCLUSION, respectively), and then an increment step (positive and negative) is varied to get the most optimal combination.

This equation combines the score (S_{c}) attributed by the original weighting (ltc) for each feature (W_{old}) found in the query with a boosting factor (k_{c}). The boosting factor was derived from the score provided by the argumentative classifier for each classified sentence. For these experiments, the parameters were determined with a tuning set (TS), one of the four query sets, and the final evaluation was done using the remaining three sets, the validation sets (VS). The document feature factor (atn) remained unchanged.



CONCLUSION|00160116| The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.

METHODS|00160119| Tumors were classified according to well-established histologic criteria to determine stage of disease; the system proposed by Robson was used.

METHODS|00162303| Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied.

cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs.

PURPOSE|00167817| Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited.

RESULTS|00155338| Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV.

Fig. 3 Classification results for the abstract shown in Fig. 1. In each box, the attributed class is first, followed by the score for the class, followed by the extracted text segment. In this example, one of RESULTS sentences is misclassified as METHODS.

4. Results

In this section, we described the generation of the baseline measure and the effects of different conditions on this baseline.

4.1. Comparison of text index parameters

The use of a domain specific thesaurus tends to improve the MAP when compared to the citation benchmark, 0.1528 versus 0.1517 for ltc.atn and 0.1452 versus 0.1433 for atc.atn (Table 3). The ltc.atn weighting schema in combination with the thesaurus produced the best results, therefore these parameters were more likely to retrieve

Table 3 Mean average precision (MAP) for each query set (1—4) with different term weighting schemas

	atc.atn	atc.atn + T	ltc.ltn	ltc.ltn + T
Set 1	0.140	0.144	0.150	0.1524
Set 2	0.1417	0.1431	0.1528	0.1534
Set 3	0.1438	0.1477	0.1431	0.1530
Set 4	0.1476	0.1465	0.1529	0.1539
Average	0.1433	0.1452	0.1517	0.1532

The last row gives the average MAP. T represents the thesaurus.

abstracts found in the citation index and thus were used for all subsequent experiments.

4.2. Argumentation-based retrieval

For demonstrating that argumentative features can improve document retrieval, we first determined which argumentative class was the most content bearing. Subsequently, we combined the four argumentative classes to again improve document retrieval.

To determine the value of each argumentative move in the retrieval, the argumentative categorizer first parses each query abstract, generating four groups each representing a unique argumentative class. The document collection was separately queried with each group. Table 4 gives the MAP measures for each type of argumentation. Table 4 shows the sentences classified as PURPOSE provide the most useful content to retrieve similar documents. Baseline precision of 62.5% is achieved when using only this section of the abstract. The CONCLUSION move is the second most valuable at 56% of the baseline. The METHODS and RESULTS sections appear less content bearing for retrieving similar documents, 16.4% and 17.6%, respectively, of the baseline. Each argumentative set represents roughly a quarter of the textual content of the original abstract. Querying with the PURPOSE section (25% of the available textual material) realizes almost 2/3 of the average precision and for the CONCLUSION section, it is more than 50% of the baseline precision. In information retrieval queries and documents are often seen as symmetrical elements. This relative symmetry suggests that argumentative moves could be used as a technique to reduce the size of the indexed document collection or to help indexing pruning in large repositories [32].

Table 4 MAP results from querying the collection using only the argumentative move

	MAP	Percent
PURPOSE	0.0958	62.5
METHODS	0.0251	16.4
RESULTS	0.0270	17.6
CONCLUSION	0.0858	56
BASELINE ltc.atn + T	0.1532	100

4.3. Argumentative overweighting

As implied in Table 4, Table 5 confirms that overweighting the features of PURPOSE and CONCLUSION sentences results in a gain in average precision (respectively +3.39% and +3.98% for CONCLUSION and PURPOSE) as measured by citation similarity. More specifically, Table 5 demonstrates the use of PURPOSE and CONCLUSION as follows:

- PURPOSE applies a boosting coefficient to features classified as PURPOSE by the argumentative classifier;
- CONCLUSION applies a boosting coefficient to features classified as CONCLUSION by the argumentative classifier;
- COMBINATION applies two different boosting coefficients to features classified as CONCLUSION and PURPOSE by the argumentative classifier.

The results, in Table 5, from boosting PURPOSE and CONCLUSION features are given alongside the MAP and show an improvement of precision at the 5- and 10-document level. At the 5-document level the advantage is with the PURPOSE features, but at the 10-document level boosting the CONCLUSION features is more effective. While the improvement brought by boosting PURPOSE and CONCLUSION features, when measured by MAP is modest (3—4%), the improvement observed by their optimal combination reached a significant improvement: +5.48%. In contrast, the various combinations to boost RESULTS and METHODS sections did not lead to any improvement. Beyond the strict quantitative results, a 10% improvement of the Precision at 10 means that out of 10 retrieved documents, a relevant one is added when argumentative boosting is applied.

Argumentation has typically been studied in relation to summarization [33]. Its impact on information retrieval is more difficult to establish although recent experiments [34] tend to confirm that argumentation is useful for information extraction, as demonstrated by the extraction of gene functions for LocusLink curation [38].[2] Similarly, using the argumentative structure of scientific articles has

[2] http://www.ncbi.nlm.nih.gov/projects/LocusLink/.

Table 5 Retrieval results for the argumentative classes PURPOSE and CONCLUSION, and the combination of both classes

	MAP	Precision at 5	Precision at 10
ltc.atn + T	0.1532 (100%)	0.2080	0.1840
PURPOSE	0.1593 (+3.98%)	0.2240	0.1760
CONCLUSION	0.1584 (+3.39%)	0.2160	0.1920
COMBINATION	0.1616 (+5.48%)	0.2320 (+11.5%)	0.1960 (+6.5%)

been proposed to reduce noise [35] in the assignment of Gene Ontology codes as investigated in the BioCreative challenge.[3] In particular, it was seen that the use of 'Material and Methods' sentences should be avoided for annotating proteins with the Gene Ontology. However, it is not clear how this result echoes with the poor importance of the METHODS section for a related article search task.

5. Conclusion and future work

We have reported on the construction of an information retrieval engine tailored to search for documents with similar citations in MEDLINE collections. The tool retrieves similar documents by giving more weight to features located in PURPOSE and CONCLUSION segments. The RESULTS and METHODS argumentative moves are reported here as less useful for such a retrieval task. Evaluated on a citation benchmark, the system significantly improves retrieval effectiveness of a standard vector-space engine. In this context, it would be interesting to investigate how argumentation can be beneficial to perform ad hoc retrieval tasks in MEDLINE [36].

Evidently using citation information to build our benchmark raises some questions. Authors may refer to other work in many ways to benefit the tone of their argument. Specifically, there are two major citation contexts, one where an article is cited negatively or contrastively and one where an article is cited positively, or the authors state that their own work originates from the cited work. In this study we have not made a distinction between these contexts but we consider this as an avenue for building better representations of the cited articles in future work. Finally, we are now exploring the use of the tool to detect inconsistencies between articles. We hope to use citation and content analysis to identify articles containing novel views so as to expose differences in the consensus of the research area's intellectual focus. The idea is to retrieve documents having key citation similarity but show some dissimilarity regarding a given argumentative category.

Finally, we have observed that citation networks in digital libraries are analogous to hyperlinks in web repositories. Consequently using web-inspired similarity measures may be beneficial for our purposes. Of particular interest in relation to argumentation, is the fact that citations networks, like web pages, are hierarchically nested graphs with argumentative moves introducing intermediate levels [37].

Acknowledgements

We would like to thank Patrick Brechbiehl for his assistance in organizing the computing environment as well as Ron Appel and Antoine Geissbühler for their support. The study has been partially supported by the SNF (EAGL Grant 3252BO-105755).

References

[1] P. Nakov, A. Schwartz, M. Hearst, Citances: Citation Sentences for Semantic Analysis of Bioscience Text, in SIGIR'04 Workshop on Search and Discovery in Bioinformatics.

[2] B. de Bruijn, J. Martin, Getting to the (c)ore of knowledge: mining biomedical literature, in: P. Ruch, R. Baud (Eds.), Int. J. Med. Inform. 67 (1—3, 4) (2002) 7—18.

[3] L. Hirschman, J.C. Park, J.I. Tsujii, L. Wong, C. Wu, Accomplishments and challenges in literature data mining for biology, Bioinformatics 18 (12) (2002) 1553—1561.

[4] B. Stapley, G. Benoir, BioBibliometrics: information retrieval and visualisation from co-occurrences of gene names in MEDLINE abstracts, Pac. Symp. Biocomp. 5 (2000) 526—537.

[5] C. Nédellec, M. Vetah, P. Bessières, Sentence filtering for information extraction in genomics, a classification problem, in: Proceedings PKDD, Springer-Verlag, Berlin, 2001, pp. 326—337.

[6] P.B. Dobrokhotov, C. Goutte, A.L. Veuthey, É. Gaussier, Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation, in: International Symposium on Molecular Biology (ISMB), 2003, pp. 91—94.

[7] S. Albert, S. Gaudan, H. Knigge, A. Raetsch, A. Delgado, B. Huhse, H. Kirsch, M. Albers, D. Rebholz-Schuhmann, M. Koegl, Computer-assisted generation of a

[3] http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html.

protein-interaction database for nuclear receptors, J. Mol. Endocrinol. 17 (8) (2003) 1555—1567.

[8] N. Collier, C. Nobata, J.I. Tsujii, Extracting the names of genes and gene products with a Hidden Markov Model, COLING (2000) 201—207.

[9] K. Humphreys, G. Demetriou, R. Gaizauskas, Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures, in: Proceedings of the Workshop on Natural Language Processing for Biology, Pac. Symp. Biocomp., 2000.

[10] H. Yu, V. Hatzivassiloglou, C. Friedman, I.H. Iossifov, A. Rzhetsky, W.J. Wilbur, A rule-based approach for automatically identifying gene and protein names in MEDLINE abstracts: a proposal, International Symposium on Molecular Biology (ISMB) 2002, 2002.

[11] K. Yamamoto, T. Kudo, A. Konagaya, Y. Matsumoto, Protein name tagging for biomedical annotation in text, in: ACL Workshop on Natural Language Processing in Biomedicine, 2003, pp. 65—72.

[12] G. Nenadic, H. Mima, I. Spasic, S. Ananiadou, J. Tsujii, Terminology-driven literature mining and knowledge acquisition in biomedicine, Int. J. Med. Inf. 67 (1—3) (2002) 33—48.

[13] A. Nazarenko, P. Zweigenbaum, B. Habert, J. Bouaud, Corpus-based extension of a terminological semantic lexicon, in: Recent Advances in Computational Terminology, John Benjamins, 2001.

[14] P. Srinivasan, D. Hristovski, Distilling conceptual connections from MeSH co-occurrences, MEDINFO 2004, San Francisco, California, Sept. 7—11, 2004.

[15] F. Ehrler, A. Geissbühler, A. Jimeno, P. Ruch, Data-poor categorization and passage retrieval for Gene Ontology Annotation in Swiss-Prot, BMC Bioinformatics 6 (Suppl. 1) (2005).

[16] L. McKnight, P. Srinivasan, Categorization of sentence types in medical abstracts, in: Proceedings of the 2003 AMIA Conference, 2003.

[17] H. White, Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists, J. Am. Soc. Inf. Sci. Technol. 54 (5) (2003) 423—434.

[18] E.C.M. Noyons, H.F. Moed, M. Luwel, A bibliometric study combining mapping and citation analysis for evaluative bibliometric purposes, J. Am. Soc. Inf. Sci. 50 (2) (1999) 115—131.

[19] R.R. Braam, H.F. Moed, A.F.J. van Raan, Mapping of science by combined co-citation and word analysis. I: Structural aspects, J. Am. Soc. Inf. Sci. 42 (4) (1991) 233—251.

[20] H.P.F. Peters, R.R. Braam, A.F.J. van Raan, Cognitive resemblance and citation relations in chemical engineering publications, J. Am. Soc. Inf. Sci. 46 (1) (1995) 9—21.

[21] S. Wu, F. Crestani, Methods for ranking information retrieval systems without relevance judgments, SAC 2003, ACM, 2003, pp. 811—816.

[22] I. Soborrof, C. Nicholas, P. Cahan, Ranking retrieval systems without relevance judgments, SIGIR (2001) 66—73.

[24] C. Orasan, Patterns in scientific abstracts, in: Proceedings of Corpus Linguistics, 2001, pp. 433—445.

[25] J. Swales, Genre Analysis: English in Academic and Research Settings, Cambridge University Press, UK, 1990.

[26] W. Hersh, S. Moy, D. Kraemer, L. Sacherek, D. Olson, More Statistical Power Needed: The OHSU TREC 2002 Interactive Track Experiments, Text Retrieval Conference (TREC) 2002, 2003.

[27] C. Buckley, E.M. Voorhees, Evaluating evaluation measure stability, ACM SIGIR (2000) 33—40.

[28] E. Mittendorf, P. Schäuble, Measuring the effects of data corruption on information retrieval, SDAIR Proceedings, 1996.

[29] G. Amati, C. van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inf. Syst. 20 (4) (2001) 357—389.

[30] J. Savoy, Data Fusion for Effective European Monolingual Information Retrieval, Cross Language Evaluation Forum (CLEF) Working Notes, 2004.

[31] P. Ruch, Using Contextual Spelling Correction to Improve Retrieval Effectiveness in Degraded Text Collections, COLING 2002, Morgan Kaufmann, 2002.

[32] D. Carmel, E. Amitay, M. Herscovici, Y. Maarek, Y. Petruschka, A. Soffer, Juru at TREC 10—Experiments with Index Pruning, Text Retrieval Conferences, 2001.

[33] S. Teufel, M. Moens, Summarizing scientific articles: experiments with relevance and rhetorical status, Comput. Linguistics 28 (4) (2002) 409—445.

[34] P. Ruch, R. Baud, A. Geissbühler, Learning-free text categorization, in: M. Dojat, E. Keravnou, P. Barahona (Eds.), AIME, LNAI 2780, Springer, 2003, pp. 199—208.

[35] E. Camon, D. Barrell, E. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, R. Apweiler, An evaluation of GO annotation retrieval for BioCreAtIvE and GOA, BMC Bioinformatics, 2005, 6 (Suppl 1).

[36] M. Kayaalp, A.R. Aronson, S.M. Humphrey, N.C. Ide, L.K. Tanabe, L.H. Smith, D. Demner, R.R. Loane, J.G. Mork, O. Bodenrieder, Methods for accurate retrieval of MEDLINE citations in functional genomics, in: Notebook of the Text Retrieval Conference (TREC), Gaithersburg, MD, 2003, pp. 175—184.

[37] K. Bharat, B. Chang, M. Rauch Henzinger, M. Ruhl, Who Links to Whom: Mining Linkage between Web Sites, International Conference on Data Mining, 2001, pp. 51—58.

[38] P. Ruch, C. Chichester, G. Cohen, G. Coray, F. Ehrler, H. Ghorbel, H. Müller, V. Pallotta, Report on the TREC 2003 Experiment: Genomic Track, Text Retrieval Conference (TREC), 2004.