A HYBRID APPROACH TOWARDS INFORMATION EXPANSION BASED ON SHALLOW AND DEEP METADATA

Tudor Groza and Siegfried Handschuh
Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

Keywords: Semantic Metadata, Information Expansion, Information Visualization, Knowledge Capturing.

Abstract: The exponential growth of the World Wide Web in the last decade, brought an explosion in the information space, with important consequences also in the area of scientific research. Lately, finding relevant work in a particular field and exploring links between relevant publications, became a cumbersome task. In this paper we propose a hybrid approach to automatic extraction of semantic metadata from scientific publications that can help to alleviate, at least partially, the above mentioned problem. We integrated the extraction mechanisms in an application targeted to early stage researchers. The application harmoniously combines the metadata extraction with information expansion and visualization for the seamless exploration of the space surrounding scientific publications.

1 INTRODUCTION

The World Wide Web has been an essential medium for the dissemination of scientific work in many fields. The significant rate at which scientific research outcomes are growing has inevitably led to substantial increases in the amount of scientific work published within journals, conferences, workshops. As an example, in the biomedical domain, the well-known MedLine [1] now hosts over 18 million articles, having a growth rate of 0.5 million articles / year, which represents around 1300 articles / day (Tsujii, 2009). This makes the process of finding relevant work in a particular field a cumbersome task, especially for an early stage researcher.

In addition, the lack of uniformity and integration of access to information can also be considered an associated issue with the information overload. Each event has its own online publishing means, and there exist no centralized hub linking the information, not even for communities sharing similar interests. These issues have motivated a variety of efforts. The Semantic Web Dog Food initiative (Möller et al., 2007) is a pioneering attempt in which a RDF-based repository was set up to host metadata about International and European Semantic Web conferences (and other Semantic Web events). This data is then served as linked open data to the community. While the initiative is getting more and more attention and participation from organizers of scientific events, it requires a considerable amount of manual effort to derive the metadata about the publications (shallow and deep, such as title, authors, references, claims, positions or arguments). Without any immediate reward or feedback, there is little incentive for authors to generate the metadata themselves. It is our belief that the adoption of semantic technologies to enable linked open data in the scientific publication domain can be much greater if an automatic metadata extraction solution would exist.

In this paper, we report on a hybrid approach toward automatic extraction of both shallow and deep metadata from scientific publications, which has been developed and evaluated. Our proposal combines document engineering with empirical and linguistic processing. The result of the extraction is an ontological representation of the publication, capturing the linear and rhetorical structure of the discourse (provenance and semantics), in addition to the usual Dublin-Core metadata terms. The metadata can then be exported and used in repositories like the Semantic Web Dog Food Server, or embedded into the publication and used for a tighter integration of personal information within the Social Semantic Desktop framework (Bernardi et al., 2008). We wrapped all these

[1] http://medline.cos.com/

Groza T. and Handschuh S. (2009).
A HYBRID APPROACH TOWARDS INFORMATION EXPANSION BASED ON SHALLOW AND DEEP METADATA.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 109-116
DOI: 10.5220/0002271001090116

features into a highly modularized application [2] that uses the extracted metadata for achieving information expansion and visualization. In addition, it emulates an information hub created on demand, that provides early stage researchers with an integrated view on multiple publication repositories.

The rest of the paper is structured as follows: we start by enumerating the application's requirements in Sect. 2, then we detail our approach in Sect. 3. Sect. 4 discusses the evaluation we carried out and before concluding in Sect. 6, we present similar research approaches in Sect. 5.

2 REQUIREMENTS

In order to support researchers with their needs of finding relevant scientific publications, we performed a study, that resulted in a list of requirements. The study featured a series of online surveys, with the broader scope of analyzing metadata usefulness and general reading habits. For the current paper, only parts of them are relevant [3]. In terms of results, we had 75 researchers answering the call, all acting within the Semantic Web community. The list of extracted generic requirements is summarized as follows:

Automatic Metadata Extraction. Although the vast majority of researchers agree with the importance and usefulness of metadata, almost none of them would spend the time to create it manually. Therefore, it is important to require as little effort as possible from authors / readers, and find viable ways to generate or extract the metadata automatically. In addition, we also target the extraction of the entire metadata space including abstract and references, as well as, deep metadata like claims, positions or arguments. As an example, 90.8% of the survey subjects consider the abstract important or crucial to read, and 85.4% consider deep metadata (e.g. claims) to improve the understanding of a paper, and that is also why 82.8% usually manually mark such deep metadata while reading.

Metadata Persistence. Having the metadata extracted, we need to allow the author / reader to make it persistent, thus providing the opportunity for its reuse. Both persistence options are easily achievable, i.e. (i) exporting the metadata, for direct usage in web repositories, or (ii) embedding the metadata into the original publication.

Metadata Usage. Obviously, extracting and storing the metadata would not be sufficient. We also need to use it. Considering our target users and their reading habits, we opted for using it to achieve information expansion and visualization. For example, 88.8% of the survey subjects would look for other publications of the same author or her co-authors.

3 IMPLEMENTATION

The analysis and refinement of the requirements list resulted into a workflow that supports the design of our application. In the following section we describe this workflow together with its composing parts.

3.1 Workflow

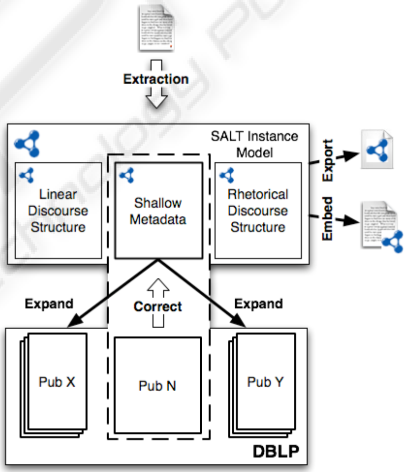


Figure 1: Application workflow.

The workflow, depicted in Fig. 1, can be split into three main parts, according to the three dimensions given by the requirements. The first part deals with the automatic extraction of metadata. Here, we currently focus only on publications published as PDF documents. The extraction module takes as input a publication and creates a SALT (Semantically Annotated LaTeX) instance model. SALT (Groza et al., 2007) represents a semantic authoring framework targeting the enrichment of scientific publications with semantic metadata. It introduces a layered model that covers the linear structure of the discourse (the Document Ontology), the rhetorical structure of the discourse (the Rhetorical Ontology) and additional annotations, e.g. the shallow metadata (the Annotation Ontology).

[2] The application can be downloaded from http://sclippy.semanticauthoring.org/

[3] The complete surveys including the results can be found at http://smile.deri.ie/surveys

In the second part of the workflow, the extracted instance model can be exported to a separate file, or embedded into the original publication (by means of the same extraction module). The third and last part uses the shallow metadata modeled by the SALT instances to perform information expansion. This is achieved based on different existing publication repositories. The current implementation realizes the expansion based on DBLP [4]. This represents only the first step, as the design of the application allows dynamic integration of different other expansion modules, thus transforming it into an information hub. In addition, the expanded information can be used to correct or improve the extracted metadata.

3.2 Extraction of Shallow Metadata and the Linear Discourse Structure

The extraction of shallow metadata and linear discourse structure currently works only on PDF publications. We have developed a set of algorithms that follow a low-level document engineering approach, by combining mining and analysis of the publication content based on its formatting style and font information. Each algorithm in the set deals with one aspect of shallow metadata. Thus, there is an authors extraction algorithm, one for extracting the abstract, one for the references and last one for the linear discourse structure.

Detailing the actual algorithms is out of the scope of the current paper. Nevertheless we will give an example of how the authors extraction works. There are four main processing steps:

1. We merge the consecutive text chunks on the first page that have the same font information and are on the same line (i.e. the Y coordinate is the same);
2. We select the text chunks between the title and the abstract and consider them author candidates
3. We linearize the author candidates based on the variations of the Y axis
4. We split the author candidates based on the variations of the X axis

To have a better picture of how the algorithm works, Fig. 2 depicts an example applied on a publication that has the authors structured on several columns. The figure shows the way in which the authors' columns containing the names and affiliations are linearized, based on the variation of the Y coordinate. The arrows in the figure show the exact linearization order. The variations on the X axis can be represented in a similar manner.

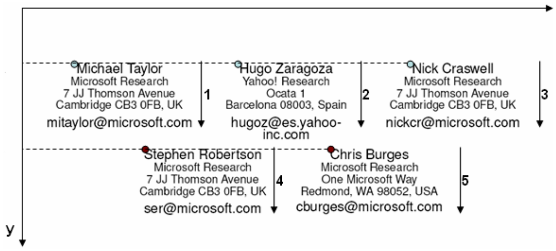


Figure 2: Authors extraction algorithm example.

3.3 Extraction of the Rhetorical Discourse Structure

For the extraction of deep metadata we followed a completely different approach. We started from adopting, as foundational background, the Rhetorical Structure of Text Theory (RST) [5]. RST was first introduced in (Mann and Thompson, 1987), with the goal of providing a descriptive theory for the organization and coherence of natural text. The theory comprises a series of elements, from which we mention the most important ones, i.e. (i) Text spans, and (ii) Schemas. *Text spans* represent uninterrupted linear intervals of text that can have the roles of Nucleus or Satellite. A Nucleus represents the core (main part / idea) of a sentence or phrase, while the Satellite represents a text span that complements the Nucleus with additional information. One the other hand, *schemas* define the structural constituency arrangements of text. They mainly provide a set of conventions that are either independent or inclusive of particular rhetorical relations that connect different *text spans*. The theory proposes a set of 23 rhetorical relations, having an almost flat structure (e.g. Circumstance, Elaboration, Antithesis, etc). In SALT, and as well as in our extraction mechanism, we adopted only a subset of 11 relations [6].

The actual extraction process comprised two phases: (i) the empirical analysis of a publication collection and development of a knowledge acquisition module, and (ii) an experiment for determining the initial probabilities for text spans to represent knowledge items (i.e. claims, positions, arguments), based on the participation in a rhetorical relation of a certain type and its block placement in the publication (i.e. abstract, introduction, conclusion or related work).

[4] http://informatik.uni-trier.de/ ley/db/

[5] RST is also the foundation of SALT

[6] See http://salt.semanticauthoring.org/ for more details on the SALT model.

In order to automatically identify text spans and the rhetorical relations that hold among them, we relied on the discourse function of cue phrases, i.e. words such as *however*, *although* and *but*. An exploratory study of such cue phrases provided us with an empirical grounding for the development of an extraction algorithm. Having as inspiration the work performed by Marcu (Marcu, 1997) we analyzed a collection of around 130 publications from the Computer Science field and identified 75 cue phrases that signal the rhetorical relations mentioned above. For each cue phrase we extracted a number of text fragments, in order to identify two types of information: (i) discourse related information (i.e. the rhetorical relations that were marked by the cue phrases and the roles of the related text spans), and (ii) algorithmic information (i.e. position in the sentence, its position according to the neighboring text spans and the surrounding punctuation). This information constitutes the empirical foundation of our algorithm that identifies the elementary unit boundaries and discourse usages of the cue phrases. The actual implementation was embedded into a GATE [7] plugin.

The second phase of the extraction process consisted of an experiment. The annotation of epistemic items (i.e. claims, positions, arguments) in a document is a highly subjective task. Different people have diverse mental representations of a given document, depending on their domain or the depth of knowledge of the document in question. Therefore, probably the most reliable annotator of a scientific publication would be its author. In order to capture the way in which people find (and maybe interpret) epistemic items, we ran an experiment. The goal of the experiment was to allow us to compute initial values for the probabilities of text spans to be epistemic items.

The setup of the experiment included ten researchers (authors of scientific publications), two collections and two tasks. The tasks represented at their basis the same task, just that each time performed on a different collection. The first collection comprised a set of ten publications chosen by us, while the second collection had 20 publications, provided by the annotators. Each annotator provided us two of her own publications. For each publication, we extracted a list of text spans (based on the presence of rhetorical relations) and presented this list to the annotators. On an average each list had around 110 items. The annotators' task was to mark in the given lists, the text spans that they considered to be epistemic items. Although the complexity of the task was high enough by its nature, we chose not to do experiment in a controlled environment. Having collected the marked lists from the annotators, we decoded the rhetorical relations hidden behind each knowledge item in the list and computed the proportional specific raw inter-annotator agreement per publication. The result was a list of proportions of specific (positive, negative and overall) raw inter-annotator agreements. Each rhetorical relation was then assigned the corresponding average of the positive agreement values, based on the originating knowledge items.

The actual extraction implemented in the system considers a fixed probability threshold for the rhetorical relations and based on the input text, it provides the list of items having the rhetorical relations probabilities above the threshold.

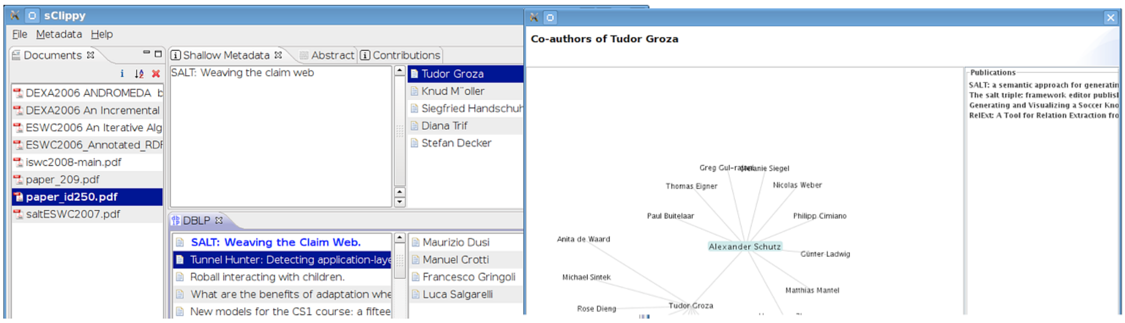
3.4 Information Expansion

The previous two steps represent together the hybrid mechanism for extracting an ontological instance model from scientific publications. The model can then be exported or embedded into the original publication, thus lifting the semantics captured in the document, into a machine-processable format. In addition, we used parts of the model to perform information expansion. The design of the application allows one to integrate multiple expansion modules, each connected to a particular publication repository. Currently, we have implemented such a module based on DBLP. On demand, the extracted shallow metadata (i.e. the title and authors) is used to search the repository for the corresponding publication. Considering that the extraction works on a best effort basis, the final metadata might contain errors both in title and in the authors' list. The user has the means for correcting it manually, or if the publication is expanded correctly, she can do it automatically.

The first element used for searching the repository is the title of the publication. In order to 'correct' (or mask) the possibly existing errors in the title, we use string similarity measures to find out the proper publication. An empirical analysis led us to using a combination of the Monge-Elkan and Soundex algorithms, with fixed thresholds. The first one analyzes fine-grained sub-string details, while the second looks at coarse-grained phonetic aspects. The publications that pass the imposed thresholds are then checked based on the existing authors. The best match is then provided to the user as a candidate for correcting the existing metadata (see the left part of Fig. 3 the title highlighted in blue). The same approach is also followed for each author of the publication.

The outcome of the expansion process features two elements: (i) a list of similar publications to the one given as input, each with its authors, and (ii) for

[7] http://gate.ac.uk/



DEXA2006 ANDROMEDA b
DEXA2006 An Incremental
ESWC2006 An Iterative Alg
ESWC2006_Annotated_RDF
iswc2008-main.pdf
paper_209.pdf
paper_id250.pdf
saltESWC2007.pdf

SALT: Weaving the claim web

Tudor Groza
Knud M¨oller
Siegfried Handschuh
Diana Trif
Stefan Decker

SALT: Weaving the Claim Web.
Tunnel Hunter: Detecting application-laye
Roball interacting with children.
What are the benefits of adaptation whe
New models for the CS1 course: a fiftee

Maurizio Dusi
Manuel Crotti
Francesco Gringoli
Luca Salgarelli

Co-authors of Tudor Groza

SALT: a semantic approach for generatin
The salt triple: framework editor publisl
Generating and Visualizing a Soccer Kno
RelExt: A Tool for Relation Extraction fro

Figure 3: Example of information expansion within our application.

each author of the given publication found, her complete list of publications existing in the respective repository. This provides the user with the chance of analyzing both publications that might have similar approaches and inspect all the publications of a particular author.

Based on the same repository, we have also implemented the option of visualizing the graph of co-authors starting from a particular author, together with their associated publications. The right side of Fig. 3 shows an example of such an exploration.

Table 1: Performance measures of the shallow metadata extraction.

	Accuracy	Prec.	Rec.	F
Title	0.95	0.96	0.98	0.96
Authors	0.90	0.92	0.96	0.93
Abstract	0.96	0.99	0.96	0.97
Sections	0.92	0.97	0.93	0.94
References	0.91	0.96	0.93	0.94

4 EVALUATION

We evaluated each of the modules described in the previous section, and performed a usability study of the overall application.

Shallow Metadata Extraction. The shallow metadata extraction was evaluated by testing the algorithms on a corpus comprising around 1200 publications, formatted with the ACM or Springer LNCS styles. The documents forming the corpus were consistent and uniform in terms of encoding and metadata content, individually for each type of formatting. As extraction from PDF documents depends on a handful of factors (e.g. encoding, encryption, etc), the evaluation results presented here, considers only the documents for which the actual extraction was valid (i.e. the PDF parser was able to read the document). Also, the results present the algorithms working on a best effort basis (no additional information is provided about the publications). Nevertheless, our next step will be to provide the means for optimizing the algorithms for particular styles and formats.

Table 1 lists the evaluation results. Overall, the title and abstract extraction algorithms performed the best, with an accuracy of 95% and respectively 96%. We observed that most of the cases in which the algorithms failed to produce a result were documents that the PDF parser managed to read but failed to actually parse. If we would eliminate this set of documents, the accuracy would probably increase with an additional 2%. The 90% accuracy of the authors extraction algorithm is mostly due to the lack of adherence to the formatting style, or presence of special symbols close to the authors' names. The sections extraction algorithm performed extremely well, with an 92% accuracy, which represents that in these cases, it managed to extract the complete tree of sections from the paper. On the other hand, the references extraction algorithm did not perform as well as we expected, and had an accuracy of only 91%.

Deep Metadata Extraction. The deep metadata extraction was tested based on a preliminary evaluation, and with an emphasis put on the extraction of the knowledge items. The setup of the evaluation was similar to the one of the experiment described in the previous section. We used two corpora (with a total of 30 publications), one with the evaluators' own papers and one containing a set of paper we chose. Each evaluator was asked to mark those text spans in the text that she considers to be knowledge items, both in her own paper and the one we assigned. In parallel, we ran our tool on the same set of publications and compiled the predicted list of candidates. At the end we computed the usual performance measures,

Table 2: Evaluation results.

Corpus	Prec.	Recall	F-Measure
I (own)	0.5	0.3	0.18
II (provided)	0.43	0.31	0.19

i.e. precision, recall and f-measure. The evaluation results are summarized in Table 2.

One could interpret of the performance measures of the extraction results in different ways. On one side, we see them as satisfactory, because they represent the effect of merely the first step from a more complex extraction mechanism we have envisioned. At the same time, if we compare them, for example, with the best precision reached by (Teufel and Moens, 2002), or 70.5, we find our 0.5 precision to be encouraging. And this is mainly because in our case there was no training involved, and we considered only two parameters in the extraction process, i.e. the rhetorical relations and the block placement, while Teufel employed a very complex naïve Bayes classifier with a pool of 16 parameters, and 20 hours of training for the empirical experiment. On the other hand, these results clearly show that we need to consider as well other parameters, such as, the presence of anaphora, a proper distinction between the different types of epistemic items, or the used verb tense, parameters which are already part of our future plans.

Secondly, the formula we have used for computing the final probabilities, within this preliminary evaluation, has a very important influence on the extraction results. Currently, we opted for a simple formula that gives more weight to the probabilities emerged from the annotation of own papers. Such an approach should be used when the automatic extraction is performed by an author on her own papers, for example, in real time while authoring them. This is clearly reflected in the positive difference in precision between the own corpus and the provided one. On the other hand, if used for information retrieval purposes, by readers and not by authors, the computation formula should be changed, so that it gives more weight to the probabilities emerged from the annotation of given papers. This practically translate into shaping the extraction results in a form closer to what a reader would expect.

Usability Study. The usability study was performed with 16 evaluators and included a series of tasks to cover all the application's functionalities. Example of tasks included: extraction and manual correction of metadata from publications, expansion of information starting from a publication or exploration of the co-author graph. At the end, the evaluators had to fill in a questionnaire, comprising 18 questions, with Likert scale or free form answers, focusing on two aspects: (i) suitability and ease of use, and (ii) design, layout and conformity to expectancies. The complete results of the questionnaire can be found at http://smile.deri.ie/sclippy-usabilitystudy.

Overall, the application scored very well in both categories. For example, the vast majority of the evaluators (on average more than 90%) found the tool both helpful and well suited for the extraction and exploration of shallow and deep metadata. In the other category, the same amount evaluators considered the application easy to learn and to use while having the design and layout both appealing and suited for the task. Possible issues we discovered in two cases. The first one would be the self-descriptiveness of the application's interface, mainly due to the lack of visual indicators and tooltips. The second case was related to the suggested list of similar publications. Although the application always proposed the exact publication selected for expansion, the rest of the list created some confusion. We believe that the cause of the confusion is the fact that the similarity measures we have adopted were not well suited.

This study led us to a series of directions for improvement. First of all, the need to make use of a more complex mechanism for suggesting similar publications. This will depend to a large extent on the repository used for expansion and on the information that it provides. For example, in the case of the ACM Portal, we will consider also text of the abstract when computing the candidates list. Secondly, augmenting the expanded information with additional elements (e.g. abstract, references, citation contexts), thus providing a deeper insight into the publications and a richer experience for the users. Lastly, the integration of the application within the Social Semantic Desktop platform. This will lead to a centralized data persistence and deeper integration and linking of the metadata into the more general context of the personal information.

5 RELATED WORK

Our approach combines different directions for achieving its goals. In the following we will try to provide a good overview of the related efforts corresponding to each research direction. We will cover mainly: (i) methods used for automatic extraction of shallow metadata, (ii) models for structuring discourse, and (iii) information visualization for scientific publications.

There have been several methods used for automatic extraction of shallow metadata, like regular expressions, rule-based parsers or machine learning. Regular expressions and rule-based systems have the advantage that they do not require any training and are straightforward to implement. Successful work has been reported in this direction, with emphasis on PostScript documents in (Shek and Yang, 2000), or considering HTML documents and use of natural language processing methods in (Yilmazel et al., 2004). A different trend in the same category is given by machine learning methods, that are more efficient, but also more expensive, due to the need of training data. Hidden Markov models (HMMs) are the most widely used among these techniques. However, HMMs are based on the assumption that features of the model they represent are not independent from each other. Thus, HMMs have difficulty exploiting regularities of a semi-structured real system. Maximum entropy based Markov models (McCallum et al., 2000) and conditional random fields (Lafferty et al., 2001) have been introduced to deal with the problem of independent features. In the same category, but following a different approach, is the work performed by Han et al. (Han et al., 2003), who uses Support Vector Machines (SVMs) for metadata extraction.

Some remarks worth to be noted here regarding a comparison between the above mentioned methods and ours. First of all, the comparison between the visual/spatial approaches (Shek's and ours) and the machine learning ones is not really appropriate. This is mainly because the latter can easily cope with general formats, while the former are "static" methods, focused on a particular format. Nevertheless, the learning methods impose a high cost due to their need of accurate training data, while the static methods have no training associated. Secondly, due to the lack of a common dataset, a direct comparison of efficiency measures cannot be realized. The main reason is because most of the machine learning methods work on plain text already extracted by some other means, while the spatial approaches work on the actual documents. Still, we consider Shek's approach to be the closest to ours, although targeting a different document format. Based on a purely empirical comparison we observe a higher accuracy for our title and authors extraction method (around 5%), as well as a higher accuracy for the linear structure extraction (around 15%), while also providing additional metadata (i.e. abstract or references).

In the area of discourse structuring, we can find a rich sphere of related work. One of the first models was introduced by (Teufel and Moens, 2002) and tried to categorize phrases from scientific publications into seven categories based on their rhetorical role. Later, the authors developed an automatic extraction approach, following a similar method to ours, starting from a corpus of manually annotated documents and a set of probabilities emerged from inter-annotator agreement studies. Teufel did not make use of any relations between the extracted items. (Shum et al., 2006) were the first to describe one of the most comprehensive models for argumentation in scientific publications, using as links between the epistemic items Cognitive Coherence Relations. They developed a series of tools for the annotation, search and visualization of scientific publications based on this model, which represent our main inspiration. The automatic extraction approach they followed was the same as the one developed by Teufel, i.e. by compiling and using a list of particular cue-phrases. Although their model is richer than the previous, due to the presence of relations, they do not make use of the placement of the item in the publication.

With respect to information visualization of scientific publications, a number of methods and tools have been reported in the literature. The 2004 InfoVis challenge had motivated the introduction of a number of visualization tools highlighting different aspects of a selected set of publications in the Information Visualization domain. (Faisal et al., 2007) reported on using the InfoVis 2004 contest dataset to visualize citation networks via multiple coordinated views. Unlike our work, these tools were based on the contents of a single file, containing manually extracted metadata. As noted by the challenge chairs, it was a difficult task to produce the metadata file (Plaisant et al., 2008) and hence the considerable efforts required, made it challenging for wide-spread use. In (Neirynck and Borner, 2007), a small scale research management tool was built to help visualizing various relationships between lab members and their respective publications. A co-authorship network visualization was built from data entered by users in which nodes represent researchers together with their publications and links show their collaborations. A similar effort to visual domain knowledge was reported by (Murray et al., 2006), with data source being bibliographic files obtained from distinguished researchers in the "network science" area. While this work was also concerned with cleansing data from noisy sources, the metadata in use was not extracted from publications themselves and no further information available from external sources such as DBLP was utilized.

6 CONCLUSIONS

This paper presented an application that integrates a series of different approaches for achieving information expansion and visualization based on extracted shallow and deep metadata. The extraction process of the shallow metadata followed a low-level document engineering approach, combining font mining and format information. On the other hand, the extraction of deep metadata (i.e. the rhetorical discourse structure) was performed based on a combined linguistic and empirical approach. The metadata was used for information expansion and visualization based on different publication repositories. The evaluation results of all of the application's components encourages us to continue our efforts in the same direction, by increasing the efficiency of the extraction mechanisms.

Future work will focus especially on improving the extraction of deep metadata by considering word co-occurrence, anaphora resolution and verb tense analysis. These improvements will also be reflected into a new iteration over the initial weights (probabilities) assigned to the epistemic items, resulted from this paper. At the application level, we will implement additional expansion modules, thus integrating more publication repositories. Also, we intend to release the application's core source code as open source so that its adoption can be directly coupled with its further development, including contributions from the researchers that would like to use it.

ACKNOWLEDGEMENTS

The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). The authors would like to thank VinhTuan Thai, Georgeta Bordea and Ioana Hulpus for their support and help.

REFERENCES

Bernardi, A., Decker, S., van Elst, L., Grimnes, G., Groza, T., Handschuh, S., Jazayeri, M., Mesnage, C., Möller, K., Reif, G., and Sintek, M. (2008). *The Social Semantic Desktop: A New Paradigm Towards Deploying the Semantic Web on the Desktop*. IGI Global.

Faisal, S., Cairns, P. A., and Blandford, A. (2007). Building for Users not for Experts: Designing a Visualization of the Literature Domain. In *Information Visualisation 2007*, pages 707–712. IEEE Computer Society.

Groza, T., Handschuh, S., Möller, K., and Decker, S. (2007). SALT - Semantically Annotated LaTeX for Scientific Publications. In *ESWC 2007*, Innsbruck, Austria.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. In *Proc. of the 3rd ACM/IEEE-CS Joint Conf. on Digital libraries*, pages 37–48.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th Int. Conf. on Machine Learning*, pages 282–289.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, Information Science Institute.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation on Natural Language Texts*. PhD thesis, University of Toronto.

McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. In *Proc. of the 17th Int. Conf. on Machine Learning*, pages 591–598.

Möller, K., Heath, T., Handschuh, S., and Domingue, J. (2007). Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects. In *Proc. of the 6th Int. Semantic Web Conference*.

Murray, C., Ke, W., and Borner, K. (2006). Mapping scientific disciplines and author expertise based on personal bibliography files. In *Information Visualisation 2006*, pages 258–263. IEEE Computer Society.

Neirynck, T. and Borner, K. (2007). Representing, analyzing, and visualizing scholarly data in support of research management. In *Information Visualisation 2007*, pages 124–129. IEEE Computer Society.

Plaisant, C., Fekete, J.-D., and Grinstein, G. (2008). Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):120–134.

Shek, E. C. and Yang, J. (2000). Knowledge-Based Metadata Extraction from PostScript Files. In *Proc. of the 5th ACM Conf. on Digital Libraries*, pages 77–84.

Shum, S. J. B., Uren, V., Li, G., Sereno, B., and Mancini, C. (2006). Modeling naturalistic argumentation in research literatures: Representation and interaction design issues. *Int. J. of Intelligent Systems*, 22(1):17–47.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28.

Tsujii, J. (2009). Refine and pathtext, which combines text mining with pathways. Keynote at Semantic Enrichment of the Scientific Literature 2009 (SESL 2009).

Yilmazel, O., Finneran, C. M., and Liddy, E. D. (2004). Metaextract: an nlp system to automatically assign metadata. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 241–242.