Unsupervised Metadata Extraction in Scientific Digital Libraries Using A-Priori Domain-Specific Knowledge

Alexander Ivanyukovich
Department of Information and Communication Technology
University of Trento
38100 Trento, Italy
Email: a.ivanyukovich@dit.unitn.it

Maurizio Marchese
Department of Information and Communication Technology
University of Trento
38100 Trento, Italy
Email: maurizio.marchese@unitn.it

Abstract— **Information extraction from unstructured sources is a crucial step in the semantic annotation of content. The challenge is in supporting an high quality automatic approach (or at least semi-automatic) in order to sustain the scalability of the semantic-enabled services of the future. Unsupervised information extraction encompasses a number of underlying research problems, such as natural language processing, heterogeneous sources integration, knowledge representation, and others that are under past and current investigation. In this paper we concentrate on the problem of unsupervised metadata extraction in the Digital Libraries domain. We propose and present a novel approach focusing on the improvement in the metadata extraction quality without involving external information sources (oracles, manually prepared databases, etc), but relying on the information present in the document itself and in its corresponding context. More specifically, we focus on quality improvements of metadata extraction from scientific papers (mainly in computer science domain) collected from various sources over the Internet. Finally, we compare the results of our approach with the state of the art in the domain and discuss future work.**

I. Introduction

The continuing expansion of the Internet has opened many new possibilities for information creation and exchange in general and in the academic world in particular: electronic publishing, digital libraries, electronic proceedings, self-publishing and more recently blogs and scientific news streaming are rapidly expanding the amount of available scholarly digital content. Recently, we have also witnessed a major shift in the landscape of scientific publishing with projects like the Open Access Initiative [1]. In fact the number of open access journals is rising steadily, and new publishing models are rapidly evolving to test new ways to increase readership and access. Such new channels for academic communications are complementing and sometimes competing with traditional authorities like journals, books and conferences proceedings. The existence of such variety and size of scholarly content as well as its increasing accessibility opens the way to the development of useful semantic-enable services (like author's profiling [2], scientometrics [1], automatic science domains

mapping [2], scientific social networks analysis etc.). However for the implementation of such semantic-aware services first the accumulated and available scholarly content need to be annotated with proper and high quality semantic information.

In the specific domain of scholarly literature, the structure of the published scientific information still follows, in the majority of the cases, a number of established communication approach and patterns, i.e. a certain number of structure information, such as title, author's list, abstract, body, references et al., are always present. This fact allows adopting existing information processing techniques for both traditional and Internet-based sources, contributing to the processing and creation of structured information within this content type.

In this paper we will focus on the intersection of Information Retrieval (IR) and Digital Libraries (DL) research domains to address the problem of quality automatic information extraction from digital scientific documents. This is a first and crucial step towards the semantic annotation of the raw digital content, in a kind of knowledge supply chain, as indicated in [3].

In describing information extraction within DL we will use the term *metadata* to refer to the *structured* information obtained from text-based documents that includes but is not limited to title, authors, affiliations, year of publication, publishing source (journal, conference, etc), publishing authority (such as ACM, IEEE, Elsevier, etc) and the list of references - each including previously mentioned work. A number of standards are available describing and categorizing bibliographic and publishing metadata: for instance Dublin Core [4] and Bib-1 Attribute Set from ANSI/NISO Z39.50-2003 (ISO 23950) [5]. In the present work we limit our investigation on metadata extraction to a significant subset of such standards. In fact, here we want to describe and evaluate our approach; extension to other instances of metadata is only quantitative and not conceptual.

The two different information sources of scientific content (traditional and Internet sources) present important differences in the approach for metadata retrieval: traditional sources are usually based on manually prepared information (from certi-

---
[1]http://www.openarchives.org/
[2]http://www.rexa.info//

fied authorities such as professional associations like ACM, IEEE and commercial publishers, such as Elsevier, Springer, etc.). In this case either all records are manually processed or processing results are manually revised. This is possible because traditional sources usually belong to single authorities with their internal standards on information storage. On the other hand Internet-based sources usually belong to large open communities (single researchers, group of researcher, institutions) and do not follow specific strict standards. For instance, an academic paper that is stored in the Digital Library of the IEEE Computer Society [3] contains the appropriate metadata to support navigation through related papers (search and sort by author, by publication date, etc). The same paper can be found on the homepage of the author or in the digital repository of the affiliated academic institution. In this case, most often the metadata is not separated from the paper, or it is not structured. It needs either extraction or separate processing.

The problem of metadata extraction in the specific context of scientific Digital Libraries can be summarized as

1) identification of logical structures within single documents (header, abstract, introduction, body, references section, etc.)
2) entity recognition (author, title, reference, etc.) within single document
3) metadata recognition within single entity.

A general assumption, in current metatada extraction techniques, is based on the fact that there is a limited number of formats to structure an academic paper and to represent references. This is particularly true in Computer Science domain where even a fewer number of formats are in active use (ACM format, IEEE format, etc). This information is particularly helpful for point (1) above, but nevertheless one can achieve low quality results because of differences in formatting due to a number of reasons such as (a) authors not following the pattern, (b) specifics of text representation in columns in PDF/PS formats, (c) text pagination, (d) presence of headnotes and footnotes, etc. Obviously similar problems could be found in (2) and (3) as well, among them human errors, text extraction technical details from PDF/PS formats and initial low quality results after step (1). As a results the overall metadata quality won't be sufficient for everyday use in popular academic literature systems like CiteSeer.IST [4], Google Scholar [5] and Windows Academic Live [6].

The main contribution of this paper is a novel method for unsupervised metadata extraction based on a-priori domain-specific knowledge. Our method does not rely on any external information sources and is solely based on the existing information in the document itself as well as in the overall set of documents currently present in a given digital archive. This includes both a-priori domain-specific information and information obtained on previous processing steps. The proposed
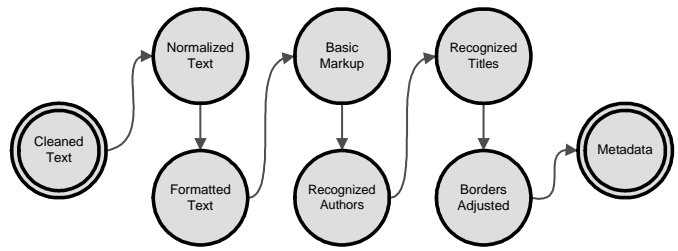
Fig. 1. Metadata Extraction Steps

method can be used in automated end-to-end information retrieval and processing systems, supporting the control and elimination of any error-prone human intervention in the process.

The remainder of the paper is organized as follows. In Section II we describe in detail the proposed approach to improve the quality of metadata extraction from scientific corpora: in particular we describe a two step procedure based on (1) pattern-based metadata extraction using Finite State Machine (FSM) and (2) statistical correction using a-priori domain-specific knowledge. In Section III we describe how the proposed approach has been applied to a large set of documents (ca. 120 K) and we provide preliminary comparison with the state of the art in the domain. In Section IV we discuss related work. Section V summarizes the results and discusses our future work.

II. THE APPROACH

The proposed approach consists of two major steps, namely

1) pattern-based metadata extraction using Finite State Machine (FSM) and
2) statistical correction using a-priori domain-specific knowledge

For the first step we have analyzed, tested and personalized for the specific application, existing state-of-the-art implementation of specialized FSM-based lexical grammar parser for fast text processing [6], [7]. For the second step we have developed and investigated statistical methods that allow metadata correction and enrichment without the need to access external information sources. In the subsequent subsection we will describe each of the two steps in details.

A. Patterns-based Metadata Extraction Using Finite State Machine

Patterns-based metadata extraction contributes to initial metadata retrieval in our approach. In contrast to the classical Information Retrieval (IR) goal, we mainly focus on the quality and not on the overall quantity of information. Core idea is in the emphasis on quality improvement within several subsequent steps, even at the cost of a limited decrease in overall metadata quantity.

This first metadata extraction step consists of a number of interim phases (see Figure 1), each implemented as a single FSM. Application of the FSM model allows simple

and formal model verification avoiding most of the human mistakes commonly involved in these tasks.

We have constructed an initial set of patterns for each grammar based on a number of small training sets (typically ca. 50 documents) from the target documents collection. Each set was manually labeled before processing and the processing results were manually evaluated. Within a limited number (ca. 10) of patterns' adjustments loops, we were able to obtain appropriate recognition quality for correct processing most of the relevant entities formats in the complete target collection (more than 120K of documents). This finding corroborated the initial assumption about the presence of a limited number of formats in a given scientific collection. According to our original idea of step-by-step quality improvement, the trade-off between metadata quality and completeness of recognition coverage on this step was shifted to the quality aspect. To this end, we allow our procedure to discard badly-formatted input, finally retaining only high-quality content and related metadata.

The major steps of metadata extraction include (see Figure 1):

1) Text normalization and special symbols removal. This covers extra spaces, new lines and tabs removal, as well as non-printable symbols handling, and references' section normalization. Moreover, it includes: text flow recognition, collateral elements detection (indexes, tables of content, pages header and footer, etc), and hyphens correction regardless of the text's language.

    These 1st-level pre-processing activities in our information extraction process, although conceptually simple, provide a number of important values that contributes to the overall quality of the subsequent information extraction. Namely:

    - Text pre-processing contribute to the more accurate textual information acquisition, i.e. correctly identified text flow (pages ordering), removal of the repeated elements that do not contribute to the structural content (headers and footers) and removal of text delimiters inside structural elements (footnotes and page numbers inside single reference, hyphens in the authors' names and titles in references, etc)
    - Text structuring contributes to the correct identification of the major structural elements within a text, i.e. Introduction section and reference to the Introduction section in a Table of content section of the article should be correctly distinguished and handled appropriately

    An extended presentation of the techniques used in this step can be found in [8].

2) Initial text tagging: separating header, abstract and references parts. This allows us to process each section separately, contributing to the improved processing speed of the subsequent FSMs.

3) References separation and initial items recognition within each single reference.

```
define program
    [head]
    [uninteresting]
    [opt references]
  | [empty]
end define

define head
    [head_begin_tag][newline]
    [opt head_line]
    [repeat other_line]
    [head_end_tag]  [newline]
end define

define head_line
    [author][repeat separator_author][delimiter]
    [repeat token_not_newline+][newline]
end define

define other_line
    [author][repeat headseparator_author][repeat trash][newline]
  | [line]
end define

define headseparator_author
    [opt space][headseparator][opt space][author]
end define
...
```

Fig. 2.   FSM: Authors recognition step

4) Authors recognition (see Figure 2) and title recognition using *invariants first* method proposed in [9]. iIn brief, this method denotes that subfields of a reference that have relatively uniform syntax, position, and composition given all previous parsing, are the first to be parsed subsequently.

5) Borders adjustments. We constructed heuristics for smart borders shift based on the number of lexical constructions from the grammar in the marked (recognized) and unmarked (not recognized) reference's region.

At present, our FSM application is context-free:, i.e. we do not compare obtained metadata with already existing ones (as partially recognized corpus or external sources). Moreover, we have design and constructed the steps in a way that grammar application is linear to the processing document's size. Both properties - context-free and linearity - contribute significantly to the overall processing speed. The use of other information (partially recognized corpus or external sources) could be used in a successive step to improve overall quality, but at the expenses of performance.

B. Statistical Correction Using A-Priori Domain-Specific Knowledge

The metadata obtained using previous step patterns can have satisfactory quality, but in general they lack in recognition coverage. For example we can have a reference with 100%-correctly recognized title, but with partially recognized authors. We still can query this metadata, but we cannot use it for the next knowledge processing level, like for instance documents clustering based on authors or documents interlinking
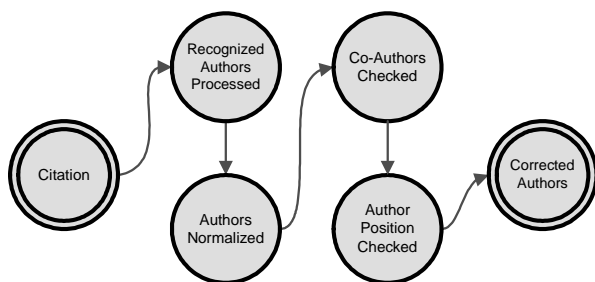
Fig. 3. Correction steps (correlation between authors)

based on their references. To tackle this issue, we developed and tested several methods for extending recognition coverage that combines (1) the partially incomplete metadata, obtained from previous step and (2) a-priori domain-specific knowledge.

To this end we have analyzed a large sample - several hundreds - of publication in computer science and extracted a limited number of usage patterns that seem to be common for the whole research domain. In particular:

- it is common to find self-citation in one author's publications - this information can be used to correct author as well as topic identification.
- it is common that in one document's reference section there are several publications with the same author - this information can be used for improving author identification (separation from other authors and from title).
- it is common to find references to the same authors within the same domain on Internet - i.e. same parent URL - (publications of the same author, home pages of authors that belong to the same organization, publications of the same institution, publications on the same event/conference). It is therefore possible to use correlation within a community for identification of authors.
- it is common that titles in references section belong to the same topic area of the paper. Therefore, it is possible to use already recognized titles for other titles identification and correction.
- it is common to find references on the same topic or number of topics within the same domain on Internet (examples are the same as those in the item before for author's identification). Also here, it is possible to use such correlation within a community for titles identification and correction.

We have use each of these assumptions for identification and correction of the corresponding metadata extraction and we have been able to statistically prove their correctness for the selected domain. For the sake of presentation of the proposed heuristic and statistical approach, we detail in the following the main procedures for (1) authors identification and correction and (2) title corrections. But the same reasoning and approach can be used for the identification and correction of other metadata present in the document, like: affiliation, keywords, type of publication (Journal, Proceedings, Workshop,...), project, event, etc.

From an operational level view, authors identification and correction can be summarized within four major steps (see Figure 3):

1) *Construction of document and community dictionaries.* All recognized authors within the same document (i.e. both paper authors and cited authors) are combined in a local dictionary (document dictionary). All recognized authors within all documents within the same domain and/or URL are combined in another dictionary (community dictionary).

2) *Co-authors dictionary building.* For each author, co-authors are checked and grouped within a separate co-author dictionary.

3) *Normalization of authors.* Authors' entries within each dictionary are normalized to the following forms: "Name Surname" and "Initials Surname". This provides a first level of disambiguation essentially using self-citation patterns. Then, we iterate the normalization step using the co-author dictionary associated to each author. This adds another level of disambiguation, using community writing patterns, within authors' initials in case of identical surname.

4) *Authors identification and correction* This last step is based on the collection of dictionaries (document, co-author and community) and aims to solve the remaining ambiguous cases using the whole knowledge present in the collection.

For the same reasons as described in previous section- i.e. simple and formal verification - authors' correction procedure was developed as FSM. Figure 4 shows a fragment of the FSM used for the implemented authors' correction procedure.

We have accomplished titles correction in a similar way (see Figure 5), however special heuristics for title borders adjustments needed to be introduced as well as a number of concepts that we use in this procedure. In particular:

- concept of "lexical formula"
- concept of "document topic"
- concept of "community topic"

Here, we define a *lexical formula* as a lexical constructs' frequency within selected logical element, i.e. the weighted set (by frequency) of words in a selected metadata field. This definition does not consider - for simplicity - any punctuation constructions and lexical constructs ordering . However, we think that on large datasets (millions of documents) this can result in better precision.

We then call "topic" a group of lexical formulas within the same document that can be merged within larger lexical formula based on the frequencies. This larger lexical formula will be referenced in the following as *document topic*.

Similarly we define *community topic* as a group of lexical formulas that can be merged together based on the frequencies of the incoming element. The difference from the document topic is that here the initial formulas belong to a set of documents originating from the same domain on Internet

```
Identify the place where new author was found:
- If the substring does not belong to any tag - mark it
- If the substring partially belongs to single Author
  tag - check for the following situations:
  - If the substring partially belongs to Author tag and Title
    tag - check for the following situations:
  - <Author>Fausto Giunchiglia Semantic</Author>
    <Title>Web</Title>
  - <Author>C. Lee </Author><Title>Giles CiteSeer.IST</Title>
    - check if <Title> is too short
    - check if last author is still OK
  - <Author>Fausto </Author> Giunchiglia. ->
    <Author>Fausto Giunchiglia</Author>
  - <Author>Fausto</Author> Giunchiglia. Semantic Web ->
    <Author>Fausto Giunchiglia</Author>. Semantic Web
  - <Author>Fausto Giunchiglia. Semantic Web</Author> ->
    <Author>Fausto Giunchiglia</Author>. Semantic Web
  - <Author>C. Lee Giles Fausto</Author> Giunchiglia. ->
    <Author>C. Lee Giles</Author>
    <Author>Fausto Giunchiglia</Author>.
  - Fausto <Author>Giunchiglia C. Lee Giles</Author> ->
    <Author>Fausto Giunchiglia</Author>
    <Author>C. Lee Giles</Author>
- If the substring partially belongs to different Author tags – check
  for situations:
- <Author>Fausto </Author>
  <Author>Giunchiglia C. Lee Giles.</Author>
- <Author>Fausto Giunchiglia C. </Author>
  <Author>Lee Giles.</Author>
...
```

Fig. 4.   FSM: Authors Correction

(typically same parent URL).

Experimentally we found out that it is sufficient to use only middle dense part of a lexical formula, omitting all top-weight constructs (usually they are represented by articles, prepositions and conjunctions) as well as low-weight constructs (usually they are represented by random constructs that are similar to random noise and do not contribute to the topic's definition).

The main operational steps for titles correction include:

1) *Recognized titles cross-check and correction.* This step includes normalized titles presentation using dictionary with weights (lexical formula) and lexical formula matching within complete collection of references located in single document.
2) *Construction of dictionaries of communities' and documents' topics.* The step includes communities and document topics identification based on the formulas from previous step and their clustering within communities.
3) *Titles' borders detection using dictionaries of topics.* This includes topics' formulas application to the references and exact borders identification based on the same patterns used for initial metadata extraction.
4) *Title-authors and title-publishing authority borders correction.* The step includes utilization of the *invariants first* method [9].
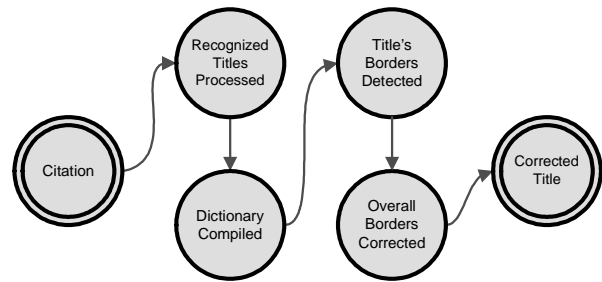

Fig. 5.   Correction steps (correlation between topics)

III. PRELIMINARY EVALUATION

Since we are working with sets of hundreds of thousands of documents, manual quality evaluation for the whole collection is not feasible. This fact has been already reported in related works [10], [11], [12] and other methodologies were proposed, based on the specifics of each concrete dataset.

In our experiments we compute quality evaluation criteria from preliminary results (extracted metadata) obtained with the proposed approach with results used in existing state-of-the-art system. In order to build a consistent reference dataset for our evaluation, we have processed results of the CiteSeer.IST Autonomous Citation Indexing (ACI) system, publicly available within its Open Archive Initiative (OAI). The results contain metadata for 570K documents collected within CiteSeer.IST project within the last several years. We have used this metadata collection as an "ideal" metadata. Here "ideal" means that this metadata was processed using external manually verified datasets (DBLP) and afterwards manually re-checked and corrected by users. Therefore, for our purpose, it is our "ideal" reference set.

We have processed list of URLs available within the "ideal" collection and were able to retrieve from the Internet around 120K of documents (other 450K documents have disappeared from their original location after they were collected by CiteSeer.IST - in 3-4 years time). In this subset of documents, we were able to process, correctly identify and extract metadata from more than 90K of documents. This set corresponds to our complete "ideal" collection used in our preliminary evaluation.

It is important to note that not all relevant metadata were present in the "ideal" collection: for instance, the references' section in each document contains only records that correspond to the documents that are present in the overall collection. Practically this means that a large part of the references are missing. We were able to overcome this limitation by retrieving missing information from the corresponding static pages within CiteSeer.IST project. This includes complete number of references and references themselves for each document, however, without clear logical constructs separation within single reference.

For our evaluation task, we have used standard quality measure criteria, namely:

$$Precision = \frac{A}{A + C} \qquad (1)$$

$$Recall = \frac{A}{A + B} \qquad (2)$$

$$F_{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (3)$$

where:

- A is the number of true positive samples predicted as positive; in our case A is the number of references that we have recognized in a document that is not more than the number of references in the corresponding document in the ideal set.
- B is the number of true positive samples predicted as negative; in our case it is the difference between number of references present in the ideal set and number of reference in corresponding document processed with our approach - but not less than zero.
- C is the number of true negative samples predicted as positive; in our case it is the difference between results obtained with our approach and the ideal set - but not less than zero.

In Table I we report the results of our evaluation tests. These preliminary results show that our procedure is capable of achieving the quality level that is comparable with the one present in the ideal set but without usage of external manually verified datasets and without human supervision (which is the case for the ideal set).

IV. RELATED WORK

The problem of unsupervised and quality metadata extraction is under intense research activity. The first successful Automated Citation Indexing (ACI) system - CiteSeer.IST - has tackled this problem using a combination of patterns-based approach (regular expressions) and manually prepared external databases (DBLP and others) [9]. This provided high-quality metadata within known references. Subsequent metadata quality improvements in CiteSeer.IST were accomplished involving human-based information corrections. Application of statistical models like Hidden Markov Model (HMM) [13] and Dual and Variable-length output Hidden Markov Model (DVHMM) [14] are reported to have nearly 90% accuracy however, the training set size used by the authors has the same magnitude as a processing corpora. Further metadata methods development turned to the Support Vector Machines (SVM) usage. Numerous experiments involving SVM have been accomplished within this task [11], [15], demonstrating high accuracy, recall and precision of the results obtained. However there were reported several problems connected with flexibility of such systems that could be the case for preventing wide method application in the real-world systems.

Natural Language Processing techniques belong to a different but widely used approach for metadata extraction. Experiments using Part-of-Speech (PoS) tagging [16], [13] have proven capable to provide sufficient accuracy,; however large manually-labeled corpora is usually required for training.

TABLE I
PRELIMINARY EVALUATION RESULTS

Precision	87,7%
Recall	88,5%
$F_{measure}$	88,1%

In this domain, a powerful example is the XIP parsing system [17] a modular, declarative and XML-empowered linguistic analyzer and annotator: the system takes XML-based documents as input, linguistically analyzes their textual content (robust parsing) and produces the set of annotations in an XML format as output. XIP robust parsing provides mechanisms for identifying Named Entity (NE) expressions, and extracting relations between words or group of words, e.g. relations between NE expressions.

Other approaches used for metadata extraction include grammar induction, hierarchical structuring and ontology-based approaches [18], [19].

Our approach aims to extend the state-of-the-art by contributing with a novel approach for metadata quality and coverage improvements. In distinction to the existing approaches, we do not use any external information repositories, while we emphasize the exploitation of the knowledge available within the available documents' collection.

V. CONCLUSIONS

In this paper we have presented a novel method for unsupervised metadata extraction based on a-priori domain-specific knowledge. The method does not rely on any external information sources and is solely based on the existing information in the document and in the document's context (set of documents). Combined with existing external knowledge sources the approach can further improve overall metadata quality and coverage. High quality automatic metadata extraction is a crucial step in order to move from linguistic entities to logical entities, relation information and logical relations and therefore to the semantic level of Digital Library usability. This, in turn, creates the opportunity for value-added services within existing and future semantic-enabled Digital Library systems.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] A. V. Raan, "Scientometrics: State-of-the-art," *Scientometrics*, vol. 38, no. 1, pp. 205–218, 2002.
[2] M. L. E.C.M. Noyons, H.F. Moed, "Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study," *Journal of the American Society for Information Science*, vol. 50, no. 2, pp. 115–131, 1999.
[3] R. Stecher, C. Niedere, P. Bouquet, and et al., "Enabling a knowledge supply chain: From content resources to ontologies," in *Proc. of the*.

[4] A. Powell, "Guidelines for implementing dublin core in xml," Dublin Core Metadata Initiative Recommendation, published at www.dublincore.org, 2003.

[5] NISO, "Information retrieval (z39.50): Application service definition and protocol specification," NISO Press, 2003.

[6] N. Kiyavitskaya, N. Zeni, J. R. Cordy, L. Mich, and J. Mylopoulos, "Semi-automatic semantic annotations for next generation information systems," in *Proceedings of the 18th Conference on Advanced Information Systems Engineering*, ser. Lecture Notes in Computer Science. Springer, 2006.

[7] J. Cordy, "Txl a language for programming language tools and applications," in *In Proceedings of 4th International Workshop on Language Descriptions, Tools and Applications*, ser. Electronic Notes inTheoretical Computer Science, vol. 110. Elsevier Science, 2004.

[8] A. Ivanyukovich and M. Marchese, "Unsupervised free-text processing and structuring in digital archives," in *Ist International Conference on Multidisciplinay Information Sciences and Technologies*, 2006, accepted for publication.

[9] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *IEEE Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[10] V. Petricek, I. J. Cox, H. Han, I. G. Councill, and C. L. Giles, "A comparison of on-line computer science citation databases," in *In Proceedings of the European Conference on Digital Libraries*, ser. Lecture Notes in Computer Science. Springer, 2005.

[11] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*. IEEE Computer Society, 2003, pp. 37–48.

[12] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the 5th ACM International Conference on Digital Libraries*. ACM, 2000.

[13] S. Cucerzan and D. Yarowsky, "Language independent, minimally supervised induction of lexical probabilities," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 270–277.

[14] A. Takasu, "Bibliographic attribute extraction from erroneous references based on a statistical model," in *Proceedings of the 2003 Joint Conference on Digital Libraries*. IEEE, 2003.

[15] Y. Hu, H. Li, Y. Cao, D. Meyerzon, and Q. Zheng, "Automatic extraction of titles from general documents using machine learning," in *Proceesings of the JCDL'05*, 2005.

[16] D. Besagni and A. Belaid, "Citation recognition for scientific publications in digital libraries," in *First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, 2004.

[17] S. Ait-Mokhtar, J. Chanod, and C. Roux, "Robustness beyond shallowness: incremental deep parsing," *Natural Language Engineering*, vol. 8, pp. 121–144, 2002.

[18] P. Cimiano, S. Handschuh, and S. Staab, "Towards the self-annotating web," in *Proceedings of the WWW2004*, 2004.

[19] M.-Y. Day, T.-H. Tsai, C.-L. Sung, C.-W. Lee, S.-H. Wu, C.-S. Ong, and W.-L. Hsu, "A knowledge-based approach to citation extraction," in *Proceedings of Information Reuse and Integration Conference, IRI-2005*, 2005.