

ScienceTreks: an Autonomous Digital Library System

Alexander Ivanyukovich¹, Maurizio Marchese¹ and Fausto Giunchiglia¹

¹ Department of Information and Communication Technology,
University of Trento,
via Sommarive, 14, 38050 Povo (TN), ITALY
{a.ivanyukovich,marchese,fausto}@dit.unitn.it

Abstract. The support for automation of the annotation process of large corpora of digital content is crucial for the success of semantic-aware services in the digital library domain. In this paper we first present and discuss an information extraction pipeline from digital document acquisition to information extraction, processing and management. Such information pipeline is divided in a number of operational steps. The realization of these steps in an unsupervised information system enables us to introduce the concept of an Autonomous Digital Library system. In the following, we describe in some detail a first prototype: the ScienceTreks¹ system. The proposed Autonomous Digital Library system can be used in automating end-to-end information retrieval and processing, supporting the control and elimination of error-prone human intervention in the process.

Keywords: Digital Libraries, Information Retrieval, Crawling, Text Processing, Metadata Extraction and Processing.

1 Introduction

At present we are experiencing exponential information growth and facing the problem of its management. Some of the main critical issues are: the management of very large repository of digital objects, the existence of many standards to encode the same information in natural language and the complexity of identification of information relevance (within the user's request, within the digital object and within a collection of digital objects). Distinct elements of the outlined problem are under investigation since long (library systems, search engines, natural language processing techniques, statistical methods of information analysis, etc). In our view, only recently the area is matured enough to shift the research attention from individual issues to a global, holistic approach to the problem, at least in specific, vertical domains. We focus on the vertical domain of scholarly/scientific content.

¹ <http://www.sciencetreks.com>

Different systems are currently available online: from commercial digital libraries (like Scopus², Web of Knowledge³, IEEEExplore⁴, ACM Digital Library⁵) to non-commercial digital libraries (CiteSeer.IST⁶, DBLP⁷) and current version of commercially-managed system that proposes a business model for academic search engines (like Google Scholar⁸ and Windows Academic Live⁹).

The existence of such variety and size of content as well as its increasing accessibility opens the way to semantic-enabled services (like unsupervised document clustering, author profiling¹⁰, scientometrics (Van Raan, 2002), science domains mapping (Noyons & Moed, 1999), scientific social networks analysis, etc.). However, the implementation of such semantic-aware services requires the annotation of the available content with high quality metadata.

The two different information sources of scientific content (traditional/journal-based and Internet sources) present important differences in the approach for metadata annotation: traditional sources are usually based on manually prepared information (from certified authorities such as professional associations like ACM, IEEE and commercial publishers, such as Elsevier, Springer, etc.). On the other hand the exponential increase of digital scientific publishing models - mentioned above - requires support for automation of all human-dependent parts of such annotation process.

In this paper we address the problem of automation of all steps in the creation of a semantically-enriched scientific digital library. We propose an "information extraction pipeline" from digital document acquisition, to format transformation and to quality automatic information extraction and annotation. An unsupervised information extraction pipeline implementation creates what we call an Autonomous Digital Library (A-DL) system. In this paper, we present an overall architecture for such a system and we describe in some detail a first prototype: the ScienceTreks system. In particular our prototype:

- support a broad range of methods for documents acquisition: from local file system repository to generic Internet crawling up to focused Internet crawling.

² <http://www.scopus.com/>

³ <http://www.isiwebofknowledge.com/>

⁴ <http://ieeexplore.ieee.org/>

⁵ <http://portal.acm.org/>

⁶ <http://citeseer.ist.psu.edu/>

⁷ <http://dblp.uni-trier.de/>

⁸ <http://scholar.google.com/>

⁹ <http://academic.live.com/>

¹⁰ <http://www.rexa.info>

- does not rely on any external information sources and is solely based on the existing information in the document itself and in the overall set of documents currently present in a given digital archive.
- provides API to support easy integration of external systems and tools in the existing “pipeline”. It is thus open to extension and potential improvements of metadata extraction and processing by other methods and tools.

The remainder of the paper is organized as follows. In Section 2 we present the overall system architecture of an Autonomous Digital Library system. In Section 3 we describe in some details the implementation of the individual information extraction pipeline steps in a prototype system: the ScienceTreks system. In Section 4 we discuss related work. Section 5 summarizes the results and discusses our future work.

2 Autonomous Digital Library System

Simplifying information gathering, processing and extraction is a challenging problem. In the traditional approach most of the real work is done by a human (“information engineer”) who possesses specific knowledge about the content and has special training in the information processing methods.

In this paper we propose and analyze an “information extraction pipeline” from digital document acquisition, to format transformation and to quality automatic information extraction and annotation. Such information extraction pipeline can be separated in a number of operational steps:

1. **Crawling:** in this step the sources of initial raw data (digital scientific documents) for the pipeline input are collected.
2. **Parsing and Harmonization:** this step covers documents format transformation (for example from PDF to text) as well as pre-processing operations.
3. **Metadata extraction:** in this step a number of sequential operation are supported: first logical structures within single documents (i.e. header, abstract, introduction, etc.) are identified; then single entity (references, etc.) within single document are recognized; finally metadata (authors, titles, publication authority, affiliations, etc.) within single entity are extracted.
4. **Metadata processing:** next step is focused to the creation of relations between the identified metadata - like for instance, the creation of the network of interlinked documents, i.e. citation graph, identification of topics, co-author’s analysis etc.
5. **Searching:** the gathered data (documents) and extracted metadata are indexed and mapped into a searchable database, to deliver fast, scalable, and reliable access with search and browse functionalities for humans as well as non-human (typically web services) users.

At the end of this process further recognition and formalization of the relevant metadata in proper semantic concepts can be performed in order to enable semantic-aware innovative services.

An Autonomous Digital Library (A-DL) system aims to an unsupervised execution of the information extraction pipeline steps outlined above. To this end, we have designed and implemented a prototype of scalable and distributed A-DL system that covers the identified digital library archive functionalities and is in the process of expansion to the semantic-based functionalities. The logical architecture of such A-DL system can be described in eight layers: (1) internal data structure, (2) information retrieval (3) parsing and harmonization, (4) metadata extraction, (5) metadata processing, (6) information management (search and retrieval), (7) application management, (8) interfaces. Layers (1), (7) and (8) represent infrastructural functionalities, layers (2), (3), (4), (5) and (6) represent the implementation of the information extraction pipeline. Each block in the presented “pipeline” is loosely coupled with the rest through common data representation scheme. It is important to note that, unlike other DL systems that provide to external systems only API for final metadata querying (Petinot, Giles, Bhatnagar, Teregowda, Han & Councill, 2004), our system architecture allows easy integration of external systems and tools in the existing “pipeline”.

Schematically, each document goes through a number of transactions covering document retrieval, text parsing and harmonization, metadata extraction, metadata processing, and indexing.

3 A-DL Modules’ Architecture Overview

The A-DL system consists of the five major modules implementing information extraction plus the internal support for data structure. Figure 1 presents a diagram of this architecture, indicating the flow of data in the subsystem. Here, we will skip the presentation of the specific implementation of the internal data structure, since it is out of the scope of the current paper. We only mention that is connected to the implementation of a distributed file system from the Apache Nutch project¹¹. In the next sub-sections we will describe each module, covering essential details.

¹¹ <http://nutch.org>

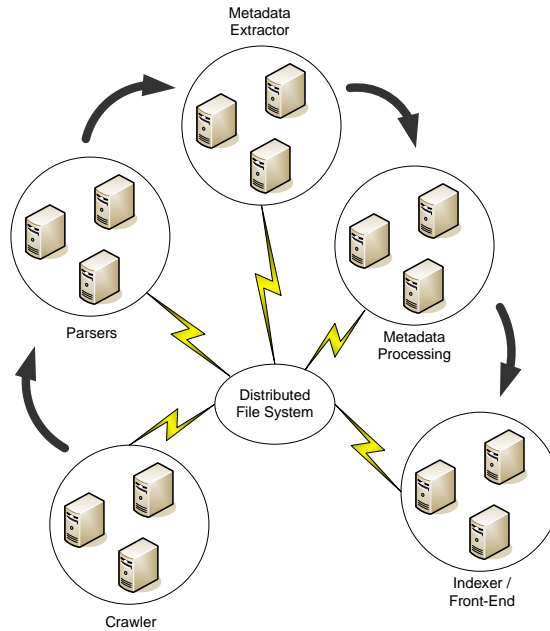


Figure 1 Main modules and dataflow in A-DL system

3.1 Crawler

Crawler module is essential for the overall system since it is the main source of initial raw data for the pipeline input. Large-scale crawler design is a relevant research problem in itself as well as technical and technological challenge (Brin & Page, 1998), (Diligenti, Coetzee, Lawrence, Giles & Gori, 2000), (Cho & Garcia-Molina, 2002). Some of the studied research issues include: crawling schemes for better coverage quality (focused crawling, random walk, etc), duplicates and near-duplicates identification (as well as content versioning), parallel crawling (independent, dynamic assignment, etc), crawler traps identification (infinite loops, generated content, etc).

Our crawler module was designed for a broad range of possible application domains so it supports several methods of documents acquisition, in particular:

- simple system bootstrapping from existing documents set, generating missing metadata if needed;
- documents retrieval from the Internet, using either a list of direct links to documents, or links to the pages with documents (1 level in-depth crawling);
- focused Internet crawling using indicated list of domains;
- Internet-wide crawling.

All discovered archived documents are uncompressed, so all eventually broken archives are discarded already on the acquisition phase. Overall crawler functionality includes: compliance with the standards (HTTP, FTP, cookies, robots.txt, etc), correct sessions handling, crawlers' traps recovery (infinite loops, etc), distributed crawling support, fault tolerance, and other minor technical features (management, monitoring, etc.).

3.2 Parser

Parser module functionality covers documents transformation to plain text and some text pre-processing operations. Our parser at the moment supports two most popular formats for publishing scientific documents: PDF and PostScript. While document to text transformation is a technical task, text pre-processing includes some research problems like text flow recognition and collateral elements detection - such as table of content, index, headers, footer, etc. - (Salton, Singhal, Mitra & Buckley, 1997) (Ivanyukovich & Marchese, 2006-1). Problem of text flow recognition is historically connected with PostScript document format¹² - articles found on the Internet can have normal and reverse page ordering. In the development of the module, we have evaluated and incorporated a number of approaches such as: numbers succession method, hyphens concatenation method, text flow prediction using Hidden Markov Model (HMM) / Dynamic Bayesian Networks (DBNs). More details on the developed methods can be found in (Ivanyukovich & Marchese, 2006-1).

3.3 Metadata Extractor

In this module first cleaned text is transformed into structured segments (abstract, introduction, references section, etc). References section is processed against individual references and afterwards individual metadata fields are extracted per reference; these include, for the moment, a subset of Dublin Core metadata standard: i.e. authors, title, conference proceedings, publication year and some other fields within document's text.

Automatic metadata extraction has been under investigation for a long time and numerous methods are available for the purpose: regular expressions, rule-based automata, machine learning, natural language processing (NLP) are among the popular ones (Han, Giles, Manavoglu, Zha, Zhang & Fox, 2003). Regular expressions and rule-based automata do not require training, are easy to implement and fast; on the other hand they require domain expert for creation and tuning, lack adaptability, their complexity increases for a moderate-to-large number of features and they are usually difficult to adapt. Machine learning techniques for information extraction include symbolic learning, inductive logic programming, grammar induction, Support Vector

¹² As well as printing process optimization.

Machines, Hidden Markov models, Dynamic Bayesian networks (DBNs) and statistical methods (Peshkin & Pfeffer, 2003). In theory, machine learning techniques are robust and adaptive, but on practice they require training set size to be the same order of magnitude as the set under investigation which limits their application. Another challenge in application of machine learning to metadata extraction is absence of false positives during training – training can be done only on true positives. NLP methods can deliver the best results but are really complex, and language-dependent and not particularly performing in term of speed. Usually they are used in a combination with other techniques.

Another research problem connected with metadata extraction is area of metadata normalization and comparison. In DL domain this includes normalization of references, normalization of authors and references comparison under uncertainty (partially overlapping information in references under comparison).

In the implementation of this module, we have followed a novel method for unsupervised metadata extraction based on a-priori domain-specific knowledge. It consists in two major steps, namely

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

For the first step we have analyzed, tested and adapted for the specific application, existing state-of-the-art implementation of specialized FSM-based lexical grammar parser for fast text processing (Kiyavitskaya, Zeni, Cordy, Mich & Mylopoulos, 2006), (Cordy, 2004). For the second step we have investigated and developed statistical methods that allow metadata correction and enrichment without the need to access external information sources. More details on the methods used can be found in (Ivanyukovich & Marchese, 2006-2).

3.4 Metadata Processor

The next step in the information extraction pipeline is dedicated to the creation of relations between metadata sets, in particular in the creation of network of interlinked documents – citation graph. This network includes bi-directional linking scheme: forward links – from a document to its references (documents that it cites) and backward links – from document to its referees (documents that cite it). Unlike other DL systems we have omitted identification documents' titles and authors on the previous metadata extraction step, because at that stage we could use only techniques described in section 3.3. On the contrary at this stage we can use documents' structure as well as metadata already collected, for more precise titles and authors recognition. At present, the approach is limited to the use of internal set of metadata. However, it can be extended to the use of existing, external high-quality metadata repositories (like DBLP and publishers' collected data

(IEEE, ACM, Elsevier, etc)). This combined approach will permit to improve the resulting citation graph quality.

3.5 Indexer and Front-end

These modules cover both typical search engine and digital library functionalities. For performance reasons we have included in the system support for index distribution over multiple PCs, fast records' location mechanisms based on Distributed File System (DFS) facilities and cache mechanisms for both queries and documents.

According to our study the text-to-binary content ratio is ~10%, the index-to-text ratio is ~30% (depends on indexing techniques – stemming usage, stop-words elimination, etc). This gives us an estimate over required memory consumption at the Front-end: for each 50Gb of processed content, we expect an addition of ~1.5Gb to the index. Index search speed is in inverse proportion to the index size . This fact adds another architectural constraint: index size should be small for usability reasons. The exact index size depends on the possible speed of read operation. Our implementation enables index distribution over a network of distributed PCs where each node can keep its part of index always in-memory thus optimizing the speed for read operations. Front-End functionalities are simple and straightforward: at present they supports metadata and full-text search, metadata retrieval, and binary content retrieval (cached versions of documents).. Additional features are limited to the citation-based ranking functionalities.

For practical reasons in our initial experiments we have used content from existing state-of-the-art DL system: we have processed results of the CiteSeer.IST Autonomous Citation Indexing (ACI) system, publicly available within the Open Archive Initiative (OAI). The results contained metadata for 570K documents collected within CiteSeer.IST project within the last years. We have processed a list of URLs available within this collection and were able to retrieve and process from the Internet around 90K of documents (450K documents have disappeared from their original location after they were collected by CiteSeer.IST - in 3-4 years time; other 30K documents were not processed by our current parser because of corrupted binaries). In this subset of documents, we were able to process - correctly identify and extract metadata - from all documents. This set corresponds to our complete collection used for preliminary evaluation of our prototype.

4. Related Work

Wide adoption of the open standards for inter-exchange in digital libraries domain like Dublin Core¹³, IEEE Learning Objects Metadata (LOM)¹⁴ and

¹³ The Dublin Core Metadata Element Set, ISO 15836: 2003;
<http://www.niso.org/international/SC4/n515.pdf>

OAI-PMH¹⁵ and recent appearance of a number of commercial digital library systems from big market players like Google¹⁶ and Microsoft¹⁷, may serve as an indicator of growth of the overall digital libraries domain. Also, existing academic DL systems are enlarging their content size. CiteSeer.IST autonomous citation indexing system (Giles, Bollacker & Lawrence, 1998) had recently reached 730K scientific articles. Specialized academic pre-prints archives like ArXiv¹⁸ in physics, Cogprints¹⁹ in cognitive science, RePEc²⁰ in economics as well as some others are in constant growth.

Most of the recent DL research activities can be summarized in a number of topics: (1) metadata description schemes and their application, (2) interoperability schemes, (3) large-scale DL systems and distributed architectures, (4) near-duplicates (revisions, corrections, etc) identification and handling and (5) semantic-enabled services application (classification, personalized DL systems, etc).

In particular, recent feasibility studies of Dublin Core, OAI-PMH and LOM metadata description standards ((Heath, McArthur, McClelland, & Vetter, 2005); (Lagoze, Krafft, Cornwell, Dushay, Eckstrom & Saylor, 2006)) have reported difficulties with standards applicability in live DL systems and can be considered as a reference for eventual standards review. The studies were based on 3-5 years of experiments and are mainly connected with high cost of deployment and maintenance. Similar studies were carried out for interoperability protocols (OAI/XOAI/ODL) between different DL systems as well as components inside single DL system ((Suleman & Fox, 2002); (Petinot, Giles, Bhatnagar, Teregowda & Han, 2004)). The studies did not report any standards shortcomings, but are rather focused on the architectural patterns in DL systems. In the scope of continuous data growth the eventual design of distributed architecture for DL systems and user requirements analysis were recently performed ((Ioannidis, Maier, Abiteboul, Buneman, Davidson, Fox, Halevy, Knoblock, Rabitti, Schek & Weikum, 2005); (Tryfonopoulos, Idreos, & Koubarakis, 2005)). The works have proposed the global evolution scheme for DL systems, and outlined existing problems such as data organization, results presentation, requests evaluation and others. Related problem of data versioning and duplicates processing was recently reviewed and new methods for near-duplicates elimination were proposed ((Yang, Callan & Shulman, 2006); (Conrad & Schriber, 2006)). Study of

¹⁴ IEEE Standard 1484.12.1, <http://ltsc.ieee.org/wg12/>

¹⁵ <http://openarchives.org>

¹⁶ Google Scholar, <http://scholar.google.com>

¹⁷ Windows Academic Live!, <http://academic.live.com>

¹⁸ ArXiv, <http://arxiv.org>

¹⁹ Cogprints, <http://cogprints.org>

²⁰ RePEc, <http://repec.org>

future evolution of DL systems applying Semantic Web methods (Kruk, Decker & Zieborak, 2005) have shown the possibility to improve user experience in content search and navigation.

Altogether these topics reflect a more global goal towards process automation in DL systems and indicate possible application areas. Our work contributes towards this goal with the proposed information extraction pipeline architecture and the corresponding implementation in a prototype of an Autonomous Digital Library system.

6 Conclusions and Future Work

In this paper we have presented and discussed an information extraction pipeline including digital document acquisition, appropriate format transformation and quality information extraction and annotation. The proposed pipeline have been implemented in a working prototype of an Autonomous Digital Library system – the ScienceTreks system – that:

- support a broad range of methods for documents acquisition: from local file system repository to generic Internet crawling up to focused Internet crawling.
- does not rely on any external information sources and is solely based on the existing information in the document itself and in the overall set of documents currently present in a given digital archive.
- provides API to support easy integration of external systems and tools in the existing “pipeline”. It is thus open to extension and potential improvements of metadata extraction and processing by other methods and tools.

Combined with existing external knowledge sources or other metadata extraction methods, 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.

References

- [1] Brin Sergey & Page Lawrence (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of the 7th WWW Conference (pp. 107-117). Elsevier Science.
- [2] Cho Junghoo & Garcia-Molina Hector (2002). Parallel Crawlers. In Proceedings of the WWW2002. ACM Press.
- [3] Conrad, Jack G. & Schriber, Cindy P. (2006). Managing déjà vu: Collection building for the identification of nonidentical duplicate

- documents. In *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 7 (pp. 921 – 932).
- [4] Cordy J., (2004), Tx1 – a language for programming language tools and applications, in *Proceedings of 4th International Workshop on Language Descriptions, Tools and Applications*, *Electronic Notes in Theoretical Computer Science*, vol. 110, Elsevier Science.
 - [5] Diligenti M., Coetzee F. M., Lawrence S., Giles C. L. & Gori M. (2000). Focused Crawling Using Context Graphs. In *Proceedings of 26th VLDB Conference*.
 - [6] Giles, C Lee, Kurt D. Bollacker & Steve Lawrence (1998). CiteSeer: An Automatic Citation Indexing System. In *Proceedings of the ACM Conference on Digital Libraries*.
 - [7] Han Hui, Giles C. Lee, Manavoglu Eren, Zha Hongyuan, Zhang Zhenyue & Fox Edward A. (2003). Automatic Document Metadata Extraction using Support Vector Machines. In *Proceedings of the 2003 Joint Conference on Digital Libraries*. IEEE.
 - [8] Heath, Barbara P., McArthur David J., McClelland, Marilyn K. & Vetter, Ronald J. (2005). Metadata lessons from the iLumina digital library. In *Communications of the ACM*, Vol. 49, No 7 (pp. 68-74). ACM Press.
 - [9] Ioannidis, Yannis, Maier, David, Abiteboul, Serge, Buneman, Peter, Davidson, Susan, Fox, Edward, Halevy, Alon Knoblock, Craig, Rabitti, Fausto, Schek, Hans & Weikum, Gerhard (2005). Digital library information-technology infrastructures. In *International Journal on Digital Libraries*, Vol. 5, No. 4 (pp. 266-274). Springer.
 - [10] Ivanyukovich Alexander & Marchese Maurizio (2006-1), Unsupervised Free-Text Processing and Structuring in Digital Archives, In *Proceedings of the InSciT2006 Conference*, Open Institute of Knowledge, Spain.
 - [11] Ivanyukovich Alexander & Marchese Maurizio (2006-2), Unsupervised Metadata Extraction in Scientific Digital Libraries Using A-Priori Domain-Specific Knowledge, submitted to 3rd Semantic Web Applications and Perspectives Workshop.
 - [12] Kiyavitskaya N., Zeni N., Cordy J.R., Mich L., and Mylopoulos J., (2006) Semi-automatic semantic annotations for next generation information systems, , in *Proceedings of the 18th Conference on Advanced Information Systems Engineering*, *Lecture Notes in Computer Science*, Springer.
 - [13] Kruk, Sebastian Ryszard, Decker Stefan & Zieborak Lech (2005). JeromeDL - Adding Semantic Web Technologies to Digital Libraries. In *Proceedings of the 16th International conference on database and expert systems applications (DEXA 2005)* (pp. 716-725). Springer.

- [14] Lagoze, Carl, Krafft, Dean, Cornwell, Tim, Dushay, Naomi, Eckstrom, Dean & Saylor, John (2006). Metadata aggregation and “automated digital libraries”: A retrospective on the NSDL experience. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (pp. 230-239). ACM Press.
- [15] Noyons L. E.C.M., Moed, H.F. (1999) Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study, (1999) *Journal of the American Society for Information Science*, vol.50, no.2, pp. 115—131.
- [16] Peshkin Leonid & Pfeffer Avi (2003). Bayesian Information Extraction Network. In Proceedings of the Eighteenth International Joint Conf. on Artificial Intelligence.
- [17] Petinot Yves, Giles C. Lee, Bhatnagar Vivek., Teregowda Pradeep B, Han Hui & Councill Isaac (2004). CiteSeer-API: Towards Seamless Resource Location and Interlinking for Digital Libraries. In Proceedings of the CIKM'04. ACM Press.
- [18] Petinot, Yves Giles C. Lee, Bhatnagar Vivek., Teregowda Pradeep B & Han Hui (2004). Enabling Interoperability For Autonomous Digital Libraries: An API To CiteSeer Services. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004). (pp. 372-373).
- [19] Salton, G., Singhal, A., Mitra, M., and Buckley, C. (1997). Automatic text structuring and summarization. *Inf. Process. Manage.* 33, 2, pp. 193-207.
- [20] Suleman, Hussein & Fox, Edward A. (2002). Designing Protocols in Support of Digital Library Componentization. In *Lecture Notes in Computer Science*, Vol. 2458 (pp. 568-582). Springer.
- [21] Tryfonopoulos Christos, Idreos Stratos, & Koubarakis Manolis (2005). LibraRing: An Architecture for Distributed Digital Libraries Based on DHTs. In Proceedings of the 9th ECDL. Springer.
- [22] Van Raan A., (2002) Scientometrics: State-of-the-art, *Scientometrics*, vol. 38, no.1, pp. 205--218, 2002.
- [23] Yang, Hui, Callan, Jamie & Shulman, Stuart (2006). Next Steps in Near-Duplicate Detection for eRulemaking. In Proceedings of the 2006 international conference on Digital government research. ACM Press.