Template mining for the extraction of citation from digital documents

Ying Ding, Gobinda Chowdhury, Schubert Foo
Division of Information Studies, School of Applied Science
Nanyang Technological University, Nanyang Avenue, Singapore 639798
(p143387632@ntu.edu.sg; asggchowdhury@ntu.edu.sg; assfoo@ntu.edu.sg)

Abstract
In order to exploit the rapid growth of digital documents in the Internet, there is an urgent need to efficiently determine and extract relevant information from these documents. This study discusses an approach to extracting citation information from digital documents. Four templates are produced by using template mining techniques, one for extracting information from citing articles and the other three for extraction information from cited articles (citations). These are subsequently applied to the chosen domain of Library and Information Science (LIS). The sub-languages of citations are examined, and the flowcharts of the four templates are described in detail. This study further describes the evaluation that was carried out manually, the results obtained and the limitations of this study. We also propose two approaches for automatically building up universal citation database: standardized template mining and universal web authoring tool based on metadata and markup language.

Keywords: template mining, cited article, citing article, information extraction, citation database.

Information Extraction and Template Mining

Information Extraction (IE) is a term that involves the activity of automatically extracting pre-specified sorts of information from short, natural language texts – typically, but by no means exclusively, newswire articles [1]. In other words, IE may be seen as the activity of populating a structured information source (or database) from an unstructured, or free text, information source. This structured database then can be used for a number of purposes: for creating a citation database, for report generation, for decision making in business, for using data-mining or artificial intelligent and neural network techniques, and so on. Most work in IE has emerged from research into rule-based systems in computational linguistics and natural language processing (NLP). IE as an area of research interest in its own right was first surveyed in Cowie and Lehnert [2]. Very broadly we can say that the field grew very rapidly from the late 1980s when DARPA, the US defense agency, funded competing research groups to pursue IE. The gathering of these research groups constituted the first of what has turned into an ongoing series of extremely productive message understanding conferences, or MUCs, that have served as key events in driving the field of IR forward [3]. The detailed review, dividing the IE researches into three groups: early work on template filing, the message understanding conferences (MUCs), and other works on information extraction, has been provided by Gaizauskas and Wilks [1].

Template mining is a particular technique used in IE. The NLP technique of template mining can be used to extract data directly from text if either the data and/or the text surrounding the data form recognizable patterns [4]. When text matches a template, the system extracts data according to instructions associated with that template. Data extraction using data mining had long been dismissed by the NLP community because of its narrow domain specificity. Areas in which template mining has been successfully applied include the extraction of proper names by Cowie and Lehnert [2]; extraction of facts from press releases related to company and financial information in systems like ARTANS [5], SCISOR [6], JASPER [7], LOLITA [8] and FIES [9]; abstracting scientific papers by Jones and Paice [10]; summarizing new product information by Shuldberg et al [11]; extraction of data from analytical chemistry papers by Postma et al [12, 13] and Postma & Kateman [14]; extraction of reaction information from experimental sections of papers in the chemistry journals by Zamora and Blower [15, 16]; processing of generic and specific chemical

designations from chemical patents by Chowdhury and Lynch [17, 18] and Kemp [19]; and extraction of bibliographic citations from full texts of patents by Lawson et al [4].

Citation Analysis and Template Mining

The problem with citation analysis in the digital age is to find a collection of Web pages or e-journals and print journal articles that are tightly enough linked, so that there is something to measure [20]. The use of citation data to locate newer works that cite important prior works in a field has long been an important technique for literature research. Unfortunately, the available indexes (e.g. SCI and SSCI) tend to be quite limited in their coverage, indexing only selected sets of print academic journals without including e-journals and Web pages. Cameron [21] called for a universal, Internet-based, bibliographic and citation database which would link every scholarly work ever written - no matter how published – to every work that it cites and every work that cites it. Such a database could revolutionize many aspects of scholarly communication: literature research, keeping current with new literature, evaluation of scholarly work, choice of publication venue, improvement of information retrieval and so on. Once the universal or semi-universal citation databases have been established automatically, bibliometric research, webmetric research will be the next era of research activities. Chowdhury [22] has mentioned that template mining can be used to develop citation databases automatically and such databases may contain information something similar to the ISI databases, such as the journal information about the citing article, author name, author address, title, abstract, keywords, as well as author, journal, title and bibliographic details of the cited articles. Lawson et al [4] have used template mining to extract citation information from the full text of chemical patents.

Harter [23] conducted a Webmetrics study with 39 scholarly journals that began electronic publication no later than 1993. This study identified the eight most highly cited e-journals. Citation and publication data for three high ranking e-journals in the study were compared to similar data for print journals in the same fields. The seven most highly cited articles from the e-journals in the study were determined.

Recently, several systems have been developed for reference links from online journal articles to other journal articles. Caplan & Arms [24] summarized the current state-of-the-art of reference linking for online journal articles. Some of the reference linking systems include: NASA Astrophysics Data System (ADS)[25], National Library of Medicine's PubMed/PubRef (PubMed) system [26], and ISI's web of science [27].

This study hypothesizes that the template mining technique can be used to extract citation information from printed and digital full-text articles so that we can establish these universal or semi-universal citation databases automatically before too long in the future. In order to prove this hypothesis, an experiment was conducted with some selected electronic journal articles, and the initial results have been quite promising. This paper provide details of this experiment and suggests some measures to be taken by the publishers and authors of electronic document in order to facilitate template mining techniques.

Methodology

The journals in this study were chosen based on the following selection criteria: (1) peer-reviewed/refereed; (2) research-oriented; (3) Full coverage of the year 1998, and (4) subject domain in LIS. Six parallel publishing journals (i.e. those journals that have both printed and electronic versions) and six pure e-journals (those that are only available in electronic forms) were chosen as the research samples. Only research-oriented journal articles were considered for investigation, since they usually have citation data. Footnotes and endnotes were excluded from this study because very small number of articles with footnotes have appeared in this study.

The parallel publishing journal sample includes *Journal of the American Society for Information, Science, Journal of Documentation, Journal of Information Science, Journal of Librarianship and Information*

Science, Program, Asian Libraries. The e-journal sample includes D-Lib magazine, Journal of electronic publishing, Katharine Sharp Review, Journal of Library Services for Distance Education, Journal of Information, Law and Technology, First Monday.

One article was selected randomly from each issue of 1998 among these 12 journals. Totally, 77 articles were selected which contain 2,179 citations. These 77 articles were separated into two sets, one (43 articles with 1,067 citations) for creating the templates and the other one (34 articles with 1,112 citations) for testing.

Templates

The objective of this study was to extract citation information (both the citing and cited references) from the digital documents. In order to create appropriate templates for this template mining study an iterative process of template specification, evaluation and modification was performed. The derivation of the templates is therefore the result of a thorough analysis of the articles selected by identifying the list of patterns associated with these articles. It is assumed that these patterns are representative of other journal collections, provided that they follow some citation patterns strictly. Experience shows that citation patterns are followed strictly in printed journals, and in fact, each printed journal has a 'house style' for citation, and authors are required to follow that, and this also goes through editorial checking and correction. However, such standards/house styles, and strict adherence to those, are not common in electronic journals.

General Citing Article Template

From each article, information on the following items was to be gathered: source journal name, volume, issue, date of publication, ISSN, title, author name, author address, author email, keywords, abstract, acknowledgment, references, about the author, and so on. Figure 1 and Figure 2 shows the flowchart of the *Citing Article Template*.

Table 1. Tokens or cue information in citing article template

INFORMATION UNIT	CUE INFORMATION (TOKENS)	PUNCTUATIONS
Journal Information	(one line or several lines) 'Vol.', 'Volume', 'Issue', 'No.', 'no.' d, d(d), 'ISSN', 'pp.', d-d;	','; '.'; space; line break; blank line;
Author name	(one line or several lines) a+ +a, c.+c.+ +a, a+ +c.+ +a	','; '&'; 'and'; Line break; blank line;
Author address	(one line or several lines) 'university', 'department', 'school', city name, country name,	','; line break; blank line;
Email	'@'	Line break
Keywords	'Keywords', 'Key words'	','; ';'; space
Abstract	paragraphs after 'Abstract' or 'Type of Article'; paragraphs before 'Introduction', 'Acknowledgements', 'Background', 'Contents'; paragraphs between 'Abstract' and 'Introduction'	Blank line
Acknowledgment	'Acknowledgements', 'Acknowledgment'	
About the author	'About the Author(s)', 'Author:', 'Author', 'Authors'	
Time	d(ignore month), 1900-1999,	
References	'References', 'References and further reading', 'Formal publications cited', 'Notes', 'Bibliography', 'ADDITIONAL READINGS'; ' Works cited'; 'Background reading'; 'A brief bibliography'; 'Notes and references';	

(Notes: **d** represents a digital number, **a** represents a String, **c** represents a character. Text that appear with '' shows that the concerned text is constant. + denotes immediately following.)

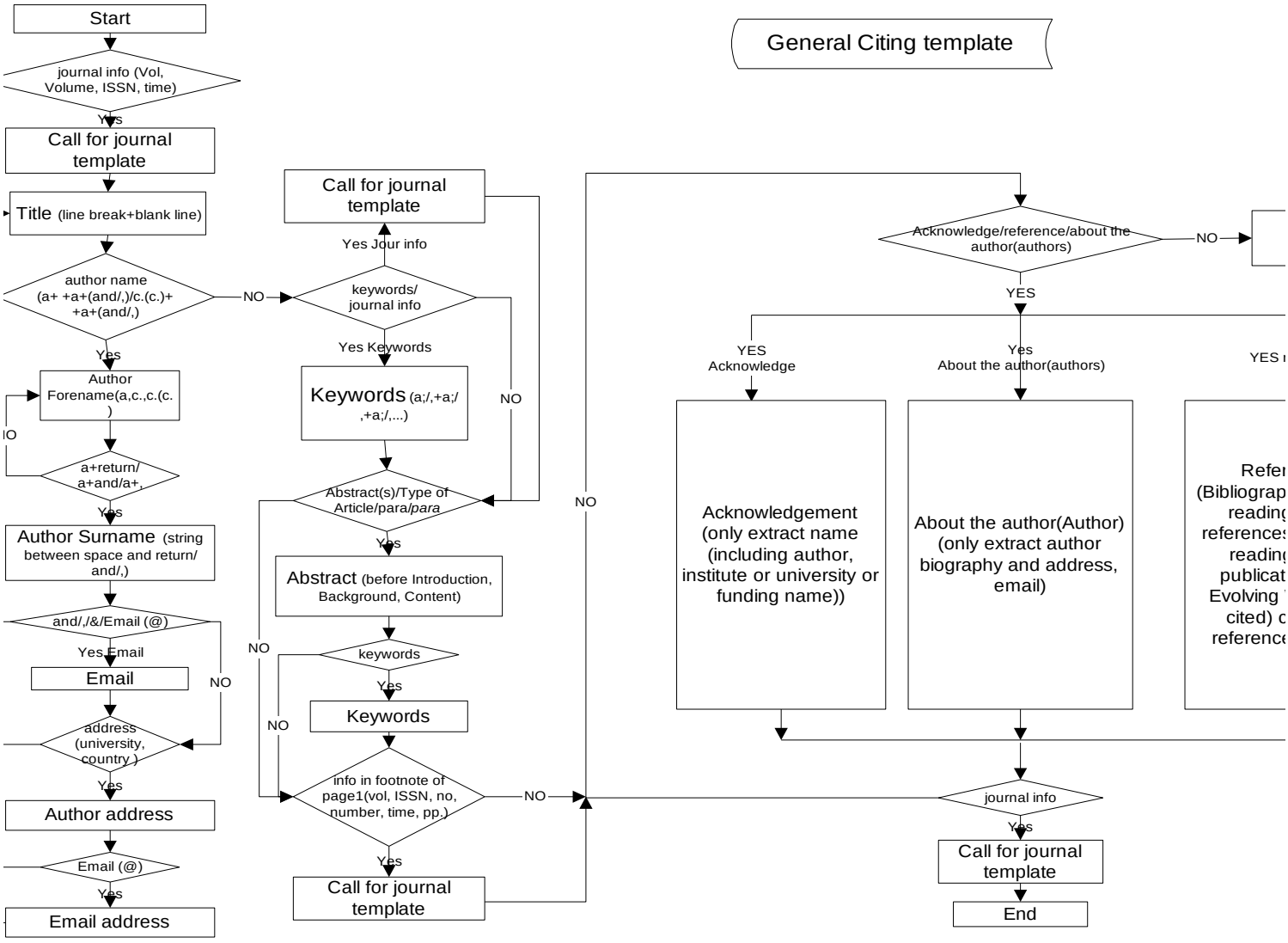

Figure 1. Flowchart of Citing Article Template

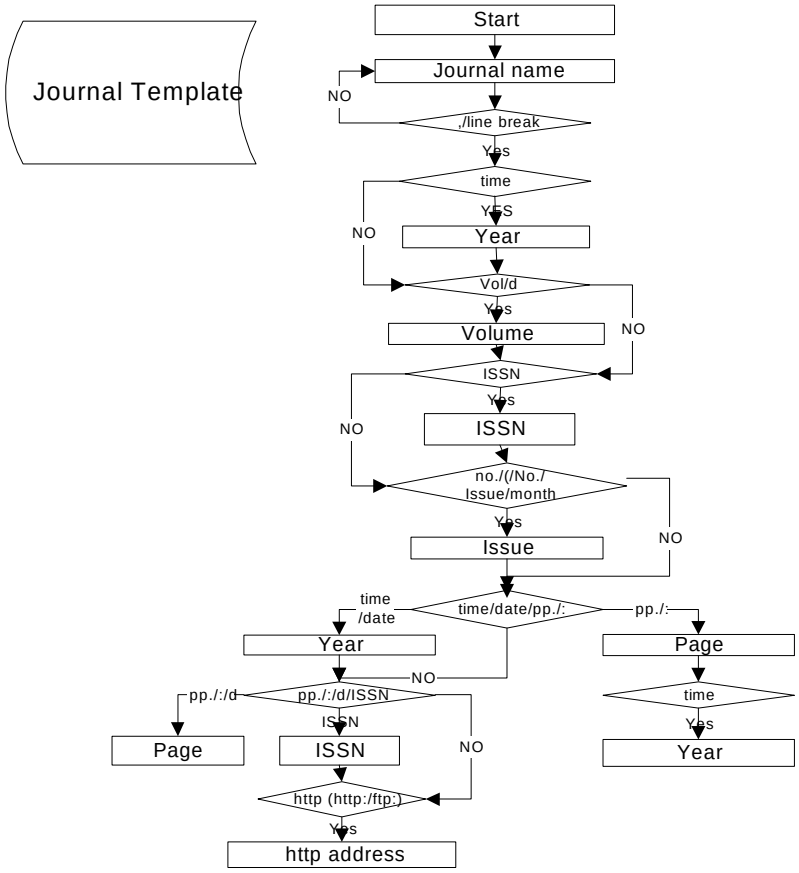

Figure 2. Flowchart of Journal Template

Each of 43 citing articles was examined to derive descriptions of the various template schemes and to identify the data to be extracted from the text by the templates. The examination also resulted in the identification of 'cue information' preceding, within or following the templates as shown in Table 1.

The formats of data elements can vary significantly. For example, author and address's position in the citing articles and even author name and address show a great deal of variations (Figure 3); so do journal information, keywords, abstract, acknowledgments, references and so on. All these add complexity of deriving the citing article templates.

General Cited Article Templates

Examination of the references indicated that there are different ways to identify the different parts of references as shown in Table 2:

- **Author name**. Although author names have several variations, they have some fixed patterns. In our study, we found two patterns for cited author name, one is author surname first, and the other one is the author forename first.
- **Position of time**. Time is easy to identify because they are four consecutive integers. But the positions of time in references are not so easy to track. It can appear at the beginning of the references (e.g. [Pinker 1994] Steven Pinker, The language instinct, New York: Harper Collins.), after the author name (e.g. Hjortgaard Christensen, F. and Ingwersen, P. (1996), "Online citation analyses: a methodological approach", Scientometrics, Vol. 37 No. 1, pp. 39-62.), at the end of the references (e.g. Mirjana Spasojevic and M. Satyanarayanan. An empirical study of a wide-area distributed file system. ACM Transactions on Computer Systems,14(2): 200-222, May 1996.) and so on. But we can use the four

consecutive digits (especially it begins with 19, later on expanding to 20) as the token to identify the time no matter where it is.

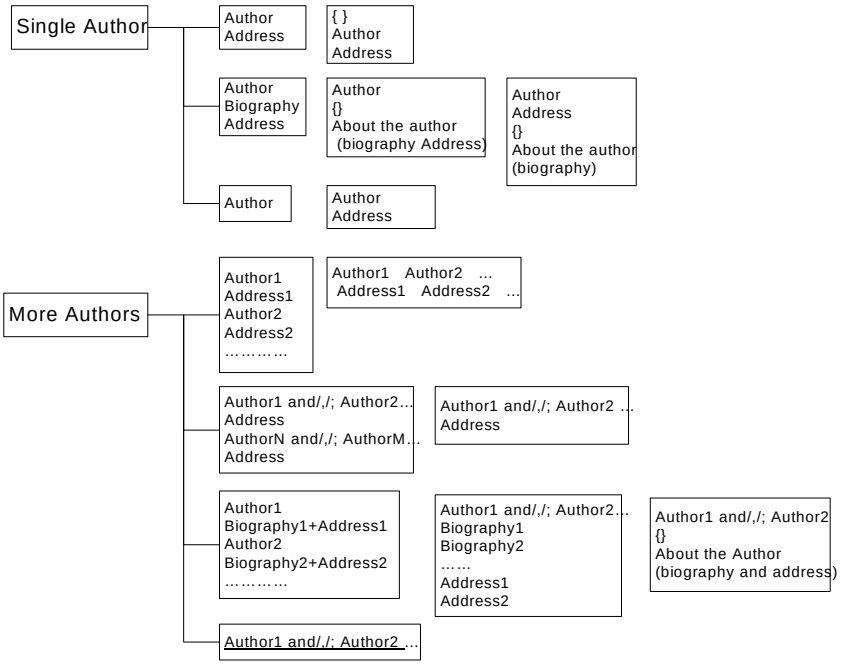

Figure 3 Variations for author and address({} represents the text body)

Table 2 Tokens or cue information in citing article template

INFORMATION UNIT	CUE INFORMATION (TOKENS)	PUNCTUATIONS
Author name	Surname,+ forename; forename+ +surname; a,+ +c.+c.; a,+ +a+ +c., a+ +a+ +a; c.+ +c.+ +a; a+ +c.+ +a;	‘,’; space, ‘;’; ‘and’; ‘&’;
Time	d(ignore month), 1900-1999,	‘(’; ‘,’; ‘.’;
Title	s+s; in; in+…+(ed.); in+…+(eds.); “s”; s+journal name; s+conference name;	‘,’; ‘”’; ‘.’; ‘?’;
Journal title	‘*Journal*’; s+number; s+vol.; s+d(d);	‘,’; ‘.’;
Book, conference, report titles and other titles	in+s; in+author name+(eds.)+s, ‘Conference’; ‘Proceedings’; ‘Seminar’; ‘Symposium’; ‘Report’; ‘Technical Report’; s+publisher information; for book includes version (such as 3rd ed, …);	‘,’; ‘”’; ‘.’;
Publisher	Publisher address+:+publishe name; publisher name+,+publisher address; publisher name; s+:+s; s+,+s; s+,; for publisher name includes ‘Inc.’; for publisher address includes CC(acronym of states, such as MA, NY..)	‘:’; ‘,’;
Page	d-d; d; ‘pp.’; ‘p.’; ‘:’;	‘pp.’; ‘p.’; ‘:’
HTTP address	‘http’; ‘ftp’; s; s; ‘Available at’; ‘Available’;	‘http’; ‘ftp’;
Journal Volume	‘Volume’; ‘Vol.’; d; d+(	‘,’; space;
Journal Issue	‘Issue’; ‘No.’; ‘no.’; (+d+);	‘(’; ‘,’;
Journal page	‘pp.’; ‘p.’; d-d; d; ‘:’	‘pp.’; ‘p.’; ‘:’;

(Notes: **d** represents a digital number, **a** represents a String, **c** represents a character.)

- **Title**. Title can include article title, book title, journal title, conference title, report title and others. The variety of titles makes extracting references by identifying titles unrealistic. Journal titles show almost as great a variation as book titles. However, some tokens and cue information hidden among them can help us to identify them, such as tokens like ‘journal’, ‘proceedings’, ‘conference’, ‘report’, ‘technical report’, ‘in’, ‘(ed.)’, ‘(eds.)’ and their abbreviations and so on. The cue information includes the punctuation, font style changing, publisher name and address, journal information (volume and page number) and so on. All these appear frequently in references and can be used as clues leading us to find the title information. Vinkler [28] observed that 55% of journal references come from only 10% of journal titles. Thus, a database containing the most frequently cited journals can be used to extract title.

It also can help us decide whether the string is journal name in citing article template. In the same way, we also can build up conference database, publisher database, country database, university database, author name database, and so on, which can make it easier to distinguish different names.

Among the various reference patterns, we generalized three cited article templates. *Cited Article Template 1* is used for the reference beginning with cited author surname, *Cited Article Template 2* is appropriate for the reference beginning with cited author forename, while *Cited Article Template 3* is appropriate for the electronic references. These are shown in Figures 4 to 6 respectively.

System Evaluation

An evaluation was carried out to study the effectiveness of these templates as a means of extracting information from new articles. 34 new articles were collected from the same journal sample as the corpus of test data. Each article was analyzed manually and information for the templates was also manually filled in.

Citing article template

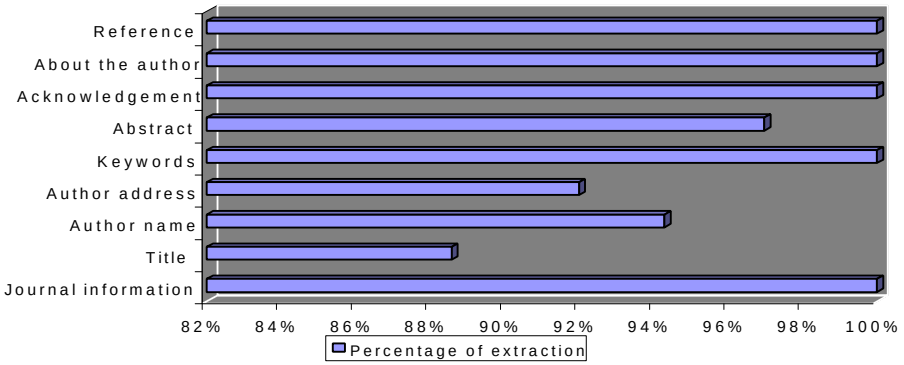

Figure 7. Distribution of information extraction from each unit in citing articles.

Figure 7 summarizes the results of evaluation for the *Citing Article Template*. Percentage of extraction is quite high in citing article template. The percentages of Reference, About the author, Acknowledgement and Journal Information are 100%. This is so since all these parts of information have fixed cue information to help the system locate them. For example, in identifying the Acknowledgement part, the system would just parse through the text to locate the token (i.e. sub heading) like 'Acknowledgement' or 'Acknowledgements', then all the information in this part will be completed for the acknowledgement slot. Thus, slots with fixed cue information or tokens will definitely have high extraction percentage. Here title slot has the lowest extraction percentage because many articles have subtitles and there are more than one blank line between title and subtitle. According to the flowchart of the *Citing Article Template*, some subtitles will be missing. The second lowest extraction percentage is author address. This is so as some author's address in these electronic journals was hyper-linked to other locations such as the author's own home page so that address information is missing. However, taken as a whole, the efficiency of the *Citing Article Template* to extract information is very promising.

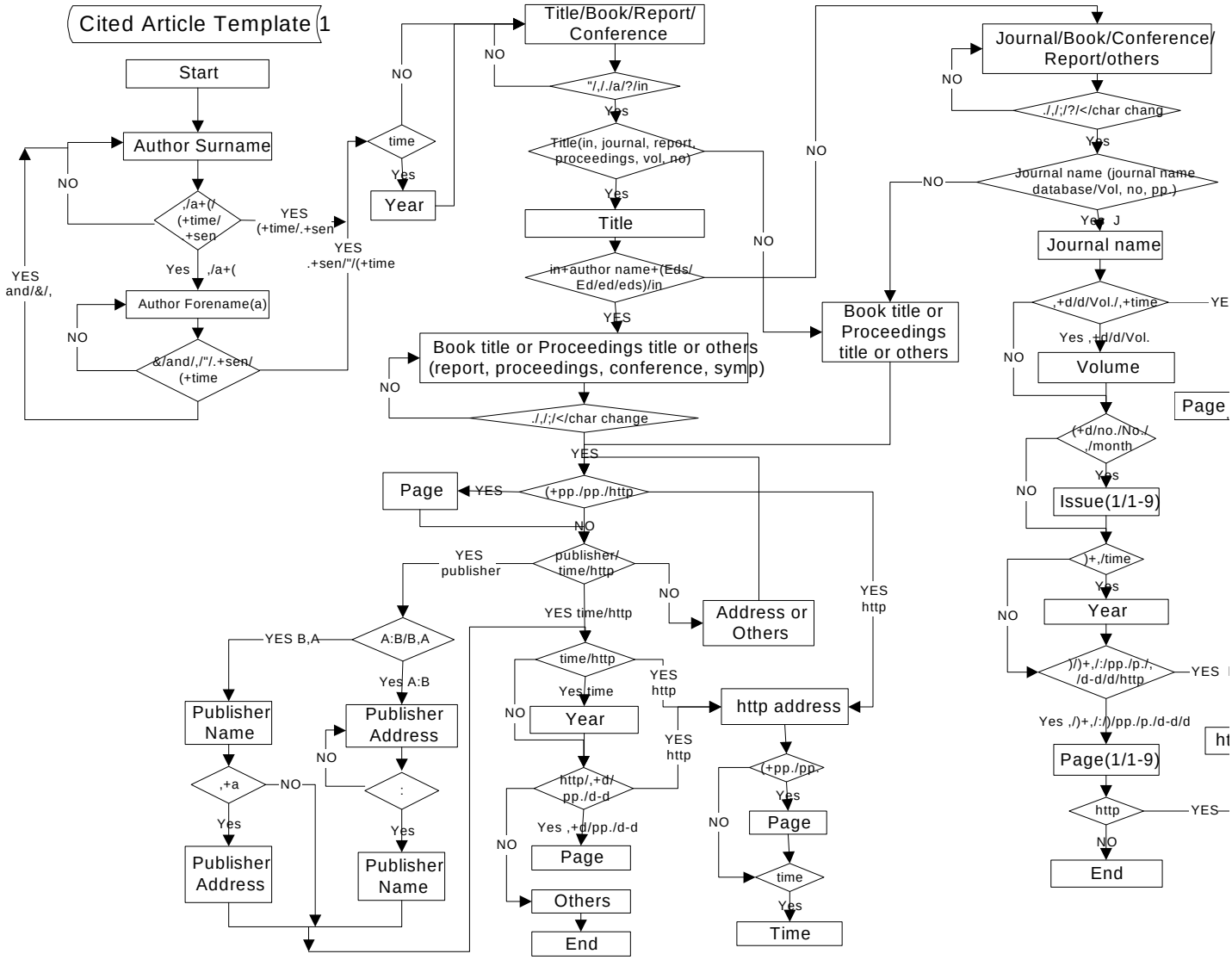

Figure 4. Flowchart of Cited Article Template 1

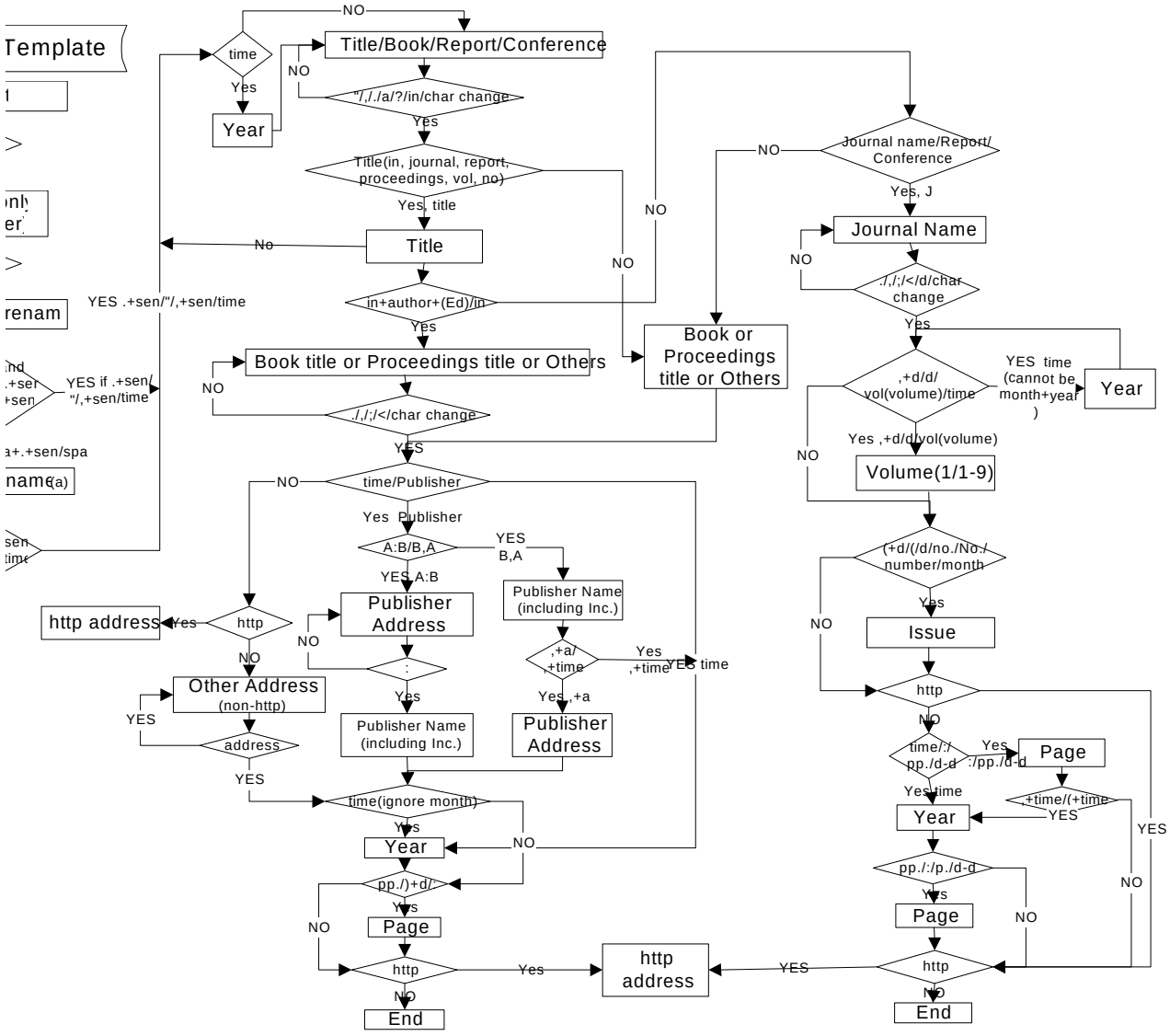

Figure 5. Flowchart of Cited Article Template 2

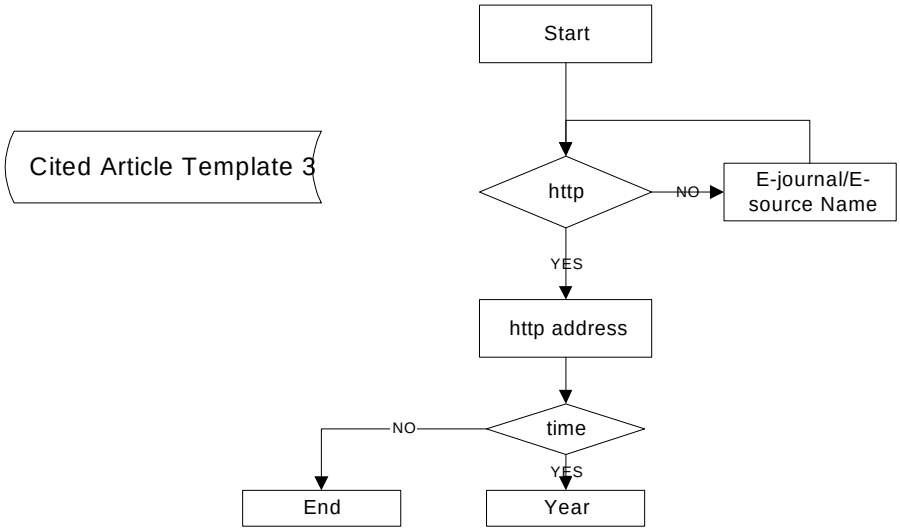

Figure 6. Flowchart of Cited Article Template 3

Cited article template

Figure 8 summarizes the results of evaluation for the cited article templates. There are much more variations in these templates in contrast to the *Citing Article Template*. It is was found that in some electronic journal articles, more than one form of citation may have been used in one same article. Thus, there is a need for more templates to cope with these larger variations. Even with these templates, many exceptions cannot be considered:

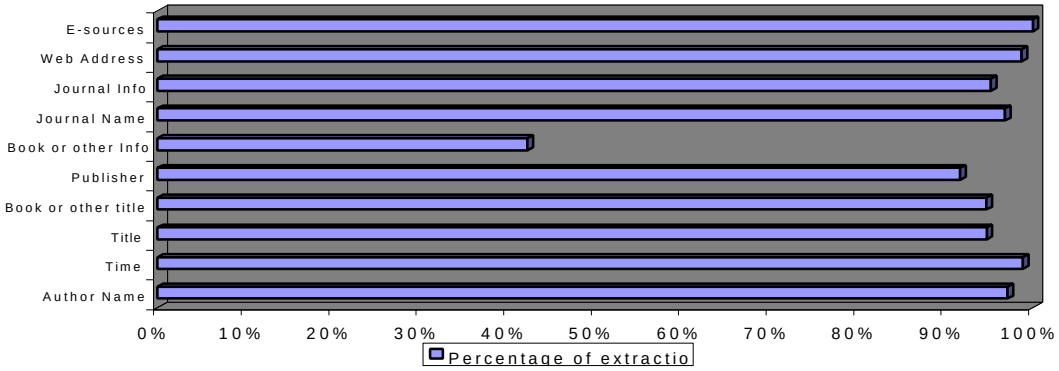

Figure 8. Distribution of information extraction from each unit in cited articles.

- In the 'Author Name' slot, Chinese authors can pose an exception since their surname does not necessarily appear last in the author name (e.g. 'Bonnie Cheuk Wai-yi', 'Gong Yitai'). Additionally, some articles do not contain the author's information. This will cause all subsequent information to be wrongly filled according to the flowchart.
- In the 'Title' slot, punctuation is the main reason that can cause exceptions. For example, in the title 'Information distributions, Part II: Resilience to ambiguity', the comma will separate the whole title into two parts, the first part been tagged to the 'Title' slot and the other tagged to 'Book or other title' Subsequent information of this cited article will also go to the wrong slots. Such forms of exceptions are very common in 'Book or other title', 'Publisher', 'Book or other Info' and 'Journal Name' slot. 'Book or other Info' slot is the one with the lowest extraction percentage. This slot not only includes book information but conference, report and other types of information that are not easily locatable or definable due to the wide variety of formats adopted by authors. Additionally, a number of citations contain more than one type of reference information. For example, a

conference citation can be subsequently followed by a book citation, or a conference citation followed by an hyperlink pointing to the electronic source.

- In the 'Journal Info' slot, information about the journal such as the volume, issue, date, page number and so on, are not always consistent and can cause exceptions. For example, 'Wired 3.07' (What does '3.07' imply?), 'Online, 20(4), 12-14, 15-18, 20-21' (Is this some form of multiple page numbers?), and so on.
- There are also special exceptions that can cause the cited article templates to fail. Examples of such text include: 'Plato, The Republic (Gyges's Ring is in Book II).', 'Carina Chocano, "Don't Worry, Be Hacky," Salon.', 'Strong pound hits Reed Elsevier, Bookseller, 15 August 1997, 8.'. Other special exceptions include those of personal communications and those that lack important parts (like author name, book name and so on) in cited articles.

Despite these exceptions, the overall result of the evaluation is quite satisfactory. Exceptions noted will provide additional information to improve these templates in future. Overall, this study has demonstrated the efficiency of template mining techniques in extraction of citing and cited reference information.

Discussion

Information extraction through template mining has widely been viewed as being closely related to NLP. In the absence of NLP capabilities, there is a need to select relevant information using one or more predictive and approximation methods. These could include statistical approaches, pattern matching, spatial positioning approach, and so on. The use of these methods is akin to obtaining some clues regarding certain words and their positions.

The proposed templates are initial prototypes that require further work in order to achieve more satisfactory results. For example, some authors put notes and references together, which means that they not only cite the articles but also add some notes to each cited article. This makes the automatic extraction of these citation formats even more complicated and challenging. Another example could be the diverse conference citation formats, electronic resource formats and other miscellaneous formats. In this study, we ignored the footnotes appearing in the text. All these difficulties and diversities point out the needs for future studies. In addition, as these templates are derived for a sample subset of available publications, further work needs to be carried out to gauge the applicability and effectiveness of the proposed templates on these other publications.

It is acknowledged that the templates' ability to recognize the citation components will be determined by the sequence and matching process found in these articles, else the extraction procedure will not reach the desired level of success. If the approach for the IE application is to be made sequence-free, then additional advanced techniques (e.g. those used in NLP processing) or heuristics will be needed to be incorporated in the approach. There are also some possibilities to improve the validity and reliability of the system. For example, building up the supporting databases, such as journal name database, publisher database, country database, and so on, will help the system improve the results. However, further research needs to be conducted for identifying the best possible approach that is both technically and economically feasible.

Although the general patterns of citation and citing articles appear to be quite simple, this study has revealed that many irregularities exist in these citations, thereby adding on to the difficulty of successful template mining. It has been observed that the frequency of irregularities is much higher in e-journals than in print journals. In an e-journal, some authors

have, even in the same article, used different ways to define the references. The reason for this could be that there are more vigorous checks in the print form by the editors and publishers in order to ensure adherence to house styles. In an e-journal, emphasis on such checks could be diluted in an attempt to bring out an online version in a shorter time, or that there is simply a lack of resource committed for online publications.

Irregularities make it difficult to detect clues to generalize the citing templates. There could be several ways to sort out these irregularities and to ensure standard style: to standardize the publishing process, to normalize different miscellaneous editorial rules and so on. One proposal by Chowdhury [22] is highly recommendable here. He proposed to prepare templates for each type of citation and to make them available online. Authors can download and make use of these templates for preparing the list of references and the whole style of article. This will ensure standard citation styles and eventually the whole structure of articles can be standardized, thereby facilitating the use of template mining with minimal or non-existent irregularities. Another extension of this proposal is to save this information using a markup language. There are two main advantages in this approach. First, it allows new citation components to be defined, generated and easily identified by the IE application. Sequencing and pattern matching of components becomes a non-issue. Second, metadata can be used to identify new information entities (e.g. ISBN or ISSN number) that need not necessarily be displayed but can be used to add a layer of value, so that new value-added applications can be subsequently built on top of the journals (e.g. to automate the filling of an interlibrary loan form).

Conclusion

This study has shown the potential of template mining to automatically create citation databases. Up to this moment in time, citation databases have largely been manually created. This is both time consuming, laborious, and prone to errors. The major hurdles to successful template mining are the irregularities that exist among the citations and the inconsistent styles among journals, especially in e-journals. Standardization is clearly a potential option to allow automation. The proposed templates can form the initial steps in deriving these 'value-added' standard templates for the future.

References

1. Gaizauskas, Robert and Wilks, Yorick Information extraction: beyond document retrieval. *Journal of Documentation*, 54, 1 (1998), 70-105.
2. Cowie, J. and Lehnert, W. Information extraction. *Communications of the ACM*, 39, 1 (1996), 80-91.
3. Sundheim, B.M. (ed.) *Proceedings of the Fourth Message Understanding Conference* [Defense Advanced Research Projects Agency] Morgan Kaufmann, Los Altos, CA, 1992.
4. Lawson, M.; Kemp, N.; Lynch, M.F. and Chowdhury, G.G. Automatic extraction of citations from the text of English language patents: An example of template mining. *Journal of Information Science*, 22, 6 (1996), 423-436.
5. Lytinen, S. L. and Gershman, A. ATRANS: automatic processing of money transfer messages. In: *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)* (1986).
6. Jacobs, P. and Rau, L. F. SCISOR: extracting information from online news. *Communications of the ACM*, 33, 11 (1990), 88—97.
7. Andersen, P.M.; Hayes, P.J.; Huettner, A.K.; Schmandt, L.M.; Nirenburg, I.B. and Weinstein, S.P. Automatic extraction of facts from press releases to generate news stories. In: *Proceedings of the 3rd Conference on Applied Natural Language Processing* (Trento, Italy, 31 march – 3 April, 1992), Morristown, NJ: Association of Computational Linguistics, pp.170—177.

8. Costantino, M.; Morgan, R.G. and Collingham, R.J. Financial information extraction using pre-defined user-definable templates in the LOLITA system. *Journal of Computing and Information Technology – CIT*, 4, 4 (1996), 241—255.
9. Chong, W. and Goh, A. FIES: Financial information extraction system. *Information Services and Use*, 17, 4 (1997), 215—223.
10. Jones, P.A. and Paice, C.D. A 'Select and generate' approach to automatic abstracting. In: T. McEnery and C.D. Paice, (Eds.), *Proceedings of the BCS 14th Information Retrieval Colloquium*, Springer-Verlag, Berlin, 1992.
11. Shuldberg, H.K; Macpherson, M.; Humphrey, P. and Corely, J. Distilling information from text: the EDS template filler system. *Journal of the American Society for Information Science*, 44, 9 (1993), 493—507.
12. Postma, G.J; van der Linden, J.R.; Smits, J.R.M. and Kateman, G. TICA: a system for the extraction of data from analytical chemical texts. *Chemometrics and Intelligent Laboratory Systems*, 9, (1990), 65—74.
13. Postma, G.J; van der Linden, J.R.; Smits, J.R.M. and Kateman, G. TICA: a system of extraction of analytical chemical information from texts. In: E.J. Karjalainen (Ed.). *Scientific Computing and Automation (Europe)*, Elsevier, Amsterdam, 1990.
14. Postma, G.J. and Kateman, G. A systematic representation of analytical chemical actions. *Journal of Chemical Information and Computer Sciences*, 33, 3 (1993), 176—181.
15. Zamora, E. and Blower, P.E. Extraction of chemical reaction information from primary journal text using computational linguistic techniques. 1. Lexical and syntactic phases. *Journal of Chemical Information and Computer Sciences*, 24, 3 (1984), 176—181.
16. Zamora, E. & Blower, P.E. (1984b). Extraction of chemical reaction information from primary journal text using computational linguistic techniques. 2. Semantic phase. *Journal of Chemical Information and Computer Sciences*, 24(3), 181—188.
17. Chowdhury, G.G. and Lynch, M.F. Automatic interpretation of the texts of chemical patent abstracts. Part 1: lexical analysis and categorization. *Journal of Chemical Information and Computer Sciences*, 32, (1992), 463—467.
18. Chowdhury, G.G. and Lynch, M.F. Automatic interpretation of the texts of chemical patent abstracts. Part 2: processing and results. *Journal of Chemical Information and Computer Sciences*, 32, (1992), 468—473.
19. Kemp, N.M. *The application of natural language processing to chemical patents*. Ph.D. Thesis. University of Sheffield, 1995.
20. Almind, T. C. and Ingwersen, P. Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation*, 53, 4 (1997), 404-426.
21. Cameron, R. D. A universal citation database as catalyst for reform in scholarly communication. *First Monday*, 2, 4 (1997). Available: http://www.firstmonday.dk/issues/issue2_4/cameron/index.html
22. Chowdhury, G.G. Template mining for information extraction from digital documents. *Library Trends*. 48(1), 1999, 181-207.
23. Harter, S. P. (1996), The impact of electronic journals on scholarly communication: A citation analysis. *The Public-Access Computer Systems Review*, 7(5). Available: http://info.lib.uh.edu/pr/v7/n5/hart7n5.html
24. Caplan, P. & Arms, W.Y. (1999). Reference linking for journal articles. *D-Lib Magazine*, 5(7/8). Available: http://www.dlib.org/dlib/july99/caplan/07caplan.html
25. NASA Astrophysics Data System. Available: http://adswww.harvard.edu.
26. The NLM PubMed Project. Available: http://www.ncbi.nlm.nih.gov/PubMed/overview.html
27. Atkins, H. (1999). The web of science. To be published in *D-Lib Magazine*, September.
28. Vinkler, P. The origin and features of information referenced in pharmaceutical patents. *Scientometrics*, 30, 1 (1994), 283-302.