Automatic Acknowledgement Indexing: Expanding the Semantics of Contribution in the CiteSeer Digital Library

Isaac G. Councill[1] **C. Lee Giles**[1,2] **Hui Han**[2] **Eren Manavoglu**[2]

[1]The School of Information Sciences and Technology

[2]Department of Computer Science and Engineering

The Pennsylvania State University, University Park, PA 16802

igc2@psu.edu {hhan,manavogl}@cse.psu.edu giles@ist.psu.edu

ABSTRACT

Acknowledgements in research publications, like citations, indicate influential contributions to scientific work; however, large-scale acknowledgement analyses have traditionally been impractical due to the high cost of manual information extraction. In this paper we describe a mixture method for automatically mining acknowledgements from research documents, using a combination of a Support Vector Machine and regular expressions. The algorithm has been implemented as a plug-in to the CiteSeer Digital Library and the extraction results have been integrated with the traditional metadata and citation index of the CiteSeer system. As a demonstration, we use CiteSeer's autonomous citation indexing (ACI) feature to measure the relative impact of acknowledged entities, and present the top twenty acknowledged entities within the archive.

Categories and Subject Descriptors

I.2.7 **[Artificial Intelligence]**: Natural Language Processing – *text analysis*.

H.3.1 **[Information Storage and Retrieval]**: Content Analysis and Indexing – *linguistic processing, indexing methods.*

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *collection*.

Keywords

Acknowledgements, Information Extraction, Text Mining, CiteSeer.

INTRODUCTION

Since the introduction of the Science Citation Index [11], researchers, funding agents, promotion and tenure committees, and others have used citation index measures to ascertain the quantity and quality of the impact of articles and authors as well as to explore the topical and social structure of scientific communities. Constructing citation indices can be a resource-intensive knowledge capture and management task. Until recently, two models for dealing with the cost of data extraction have been proposed for citations: a centralized model in which an organization pays employees for manual indexing and offers the results as a service (this model is used by The Institute for Scientific Information), and a distributed model that would shift the labor of citation indexing to authors [2,6]. Although distributed models promise to reduce the cost of indexing while increasing coverage, such systems have not been realized.

A third alternative for creating citation indices is the development of automatic, intelligent systems for citation extraction and management, as demonstrated in the CiteSeer Digital Library [17] and Google Scholar[1]. Such techniques drastically reduce the cost of data extraction and relationship identification at the expense of some lost accuracy in the data. Both CiteSeer and Google Scholar have shown that automated citation management systems can be successful, and are widely used sources of citation information.

Despite the usefulness of citations for the assessment of research contributions and network analysis, citations alone fall short of describing the full network of influence underlying primary scientific communication. In addition to referencing published material, many researchers choose to document their appreciation of important contributions through acknowledgements. Acknowledgements may be made for a number of reasons, but often imply significant intellectual debt to fellow researchers. Many funders of scientific work require acknowledgements in scientific publications, thus acknowledgement can be used to identify relationships of government and corporate entities to researchers and research output.

Acknowledgements embody a wide range of relationships among people, agencies, institutions, and research.

K-CAP'05, October 2-5, 2005, Banff, Canada.

[1] http://scholar.google.com

Classification schemes [5] have been proposed for six categories of acknowledgement: 1) moral support, 2) financial support, 3) editorial support 4) presentational support (e.g. presenting a paper at a conference), 5) instrumental/technical support, and 6) conceptual support, or peer interactive communication (PIC). Of all the categories, PIC has been considered the most important for identifying intellectual debt [18]; some researchers have considered acknowledgements of PIC to be at least as valuable as citations [9,4]. Acknowledgements of financial and instrumental support are valuable indicators of the contributions that organizations have made to research activities.

Despite their promise as an analytic tool, acknowledgements have remained a largely untapped resource. Presumably, the reason that acknowledgements are not currently included in major scientific indices has to do with cost of manual information extraction associated with traditional citation indexing. In order to make large-scale acknowledgement analyses practical it is necessary to take a similar approach to automatic citation indexing and develop automatic methods for extracting acknowledgements from documents. Niche digital libraries such as CiteSeer [17] and SMEALSearch[2] create ideal testing grounds for new document information extraction algorithms, and also provide opportunities to produce field-specific analyses of trends across scientific communities.

We have developed an algorithm for automatic acknowledgement extraction in order to extend the native knowledge capture capabilities of CiteSeer. This initial algorithm identifies acknowledging text passages and extracts the names of acknowledged entities. This data is stored in an auxiliary index alongside CiteSeer's traditional indices and special bridges have been created to integrate the new acknowledgement data into the CiteSeer system. This integration represents a structural shift in entity relationship handling in CiteSeer, requiring an extension of the traditional author-document and document-document relationships to reified relations between entities and documents, flexibly modeling the roles of entities within the data [10].

This work is the first effort to extend the semantics of research contribution beyond authorship and citations in CiteSeer. Specifically, including acknowledgements in the CiteSeer archive has the following advantages:

- *Contributing entity identification.* Governmental and corporate sponsors of research are invisible in the traditional CiteSeer metadata. Additionally, researchers who contribute to papers through informal channels are not currently identified. Identifying and tagging acknowledged entities opens CiteSeer to new uses where understanding the context of research support is desired.
- *Improved impact assessment.* By combining acknowledgement information with a citation index, the value of acknowledgements, or impact of acknowledged entities such as researchers or funders, can be measured according to the citation rank of the acknowledging paper.
- *Social network extension.* Acknowledgement information provides more entity relationships for social network analysis, including a more complete picture of scientific communication and influence.
- *Data linking.* Combining acknowledgements with entity tagging in other contexts such as authorship and affiliations allows the identification of multiple roles that unique entities may have within the scientific community.

The problem of extracting acknowledgements from research articles can be viewed as a specific case of document metadata extraction. Several approaches have been proposed for automatic metadata extraction, with the most common tools including regular expressions, rule-based parsers and machine learning algorithms. Regular expressions and rule-based parsers are easily implemented and can perform acceptably well if data are well-behaved. However, machine learning techniques are generally more robust and easily adaptable to new data. Machine learning methods used for information extraction include inductive logic programming, grammar induction, symbolic learning, hidden Markov models (HMM), and Support Vector Machines (SVM). Due to recent success using SVMs for learning in high-dimensional feature spaces [15,8], SVMs are becoming increasingly popular tools for classification. Recent work has shown it possible to recast the problem of information extraction as a classification task [3,13], and SVMs have been proven to be effective for chunk identification and named entity extraction [19,16,23].

While highly effective at metadata extraction, much recent work using machine learning for information extraction [13,22] exploits the semi-structured format of document headers for chunk identification and classification. The problem of basic acknowledgement extraction involves the identification of chunks of a single class found most often within free text. Machine learning techniques have been employed with great success for named entity extraction from unstructured text [1,20], particularly in the context of the Message Understanding Conferences. Despite the success of sophisticated tools, we have found that regular expressions work acceptably well for identifying acknowledgements within identifiable acknowledgement passages, and we have chosen to implement name extraction using regular expressions for our initial algorithm. However, while most acknowledgements are contained within identifiable passages, a significant portion of acknowledgements is found within unlabeled passages. For the more difficult case of identifying acknowledgements in unlabeled text passages, we use a SVM for passage identification.

[2] http://smealsearch.psu.edu

In this paper we discuss a mixture method for automatic acknowledgements extraction using a combination of regular expressions and a SVM for acknowledgement passage identification, and regular expressions for entity name extraction. The extraction performance analyses and results are based on experiments using computer science research papers.

As a demonstration, this paper also discusses an application of the acknowledgement extraction algorithm to documents within the CiteSeer digital library. The acknowledgements received by various entities are counted and the results are cross-referenced with citation information within CiteSeer in order to provide a relative measure of the impact each acknowledged entity has had within the archive. A prototype web interface for browsing the results is introduced. Although the extraction results presented here are limited and demonstrational, a large-scale acknowledgement analysis has recently been presented elsewhere based on results from the algorithm described in this paper [12].

The remainder of the paper is organized as follows: the data sets used for developing and training the extraction algorithm components are described, and then the acknowledgement extraction algorithm is presented in detail. The experimental results of the extraction algorithm on the test data will be presented. Finally, the preliminary integration of the acknowledgement algorithm with CiteSeer will be discussed and an application of the algorithm to documents in the CiteSeer digital library is presented.

PROBLEM AND DATASETS

Most acknowledgements in research papers are found in clearly identifiable acknowledgement sections within documents. Acknowledgement sections are easily identified using regular expressions. However, acknowledgment passages may also be found in unmarked sections, within the document header, or within footnotes. These acknowledgement passages are typically found at the beginning of documents (before the abstract or introduction, or on the first page) and at the end (before the references or first appendix). In order to identify these passages, we extract roughly the first page of the document and the last page before the reference section or the first appendix, which ever comes first. We then classify the lines of extracted text using a SVM to identify those lines containing acknowledgements.

Collections of documents and text passages containing acknowledgements were obtained from CiteSeer. Specifically, four sets of documents were obtained: 1) 800 documents containing acknowledgements in any location, 2) 400 documents containing acknowledgement sections, 3) 400 documents containing acknowledgements outside of acknowledgement sections, and 4) 200 documents containing no acknowledgements. Each dataset was created by randomly selecting documents from the archive and manually reviewing each paper. Dataset 1 was generated by creating a random sequence of document ids and manually examining the documents in order until 800 documents with acknowledgements were found. The set was used to calculate the proportion of acknowledgement passages contained within acknowledgement sections. Therefore, documents were tagged according to the type of acknowledgement passage contained. Individual acknowledgement passages within datasets 2 and 3 were identified, such that entire acknowledgement sections within dataset 2 were tagged as such, and all lines containing acknowledging text in dataset 3 were tagged (see Figure 1).

1. *represented by Anderson is right, then what remains to be done is to work out the specifics for each particular family of affixes. Otherwise, the issue of affix versus*
2. *\Lambda University of Groningen, University of Newcastle*
3. *<acknowledgement>y Many thanks to <entity>Dora Alexopoulou</entity>, <entity>Jonathan Ginzburg</entity>, <entity>Aaron Halpern</entity>, <entity>Jack Hoeksema</entity>,*
4. *<entity>Ineke Mennen</entity>, <entity>John Nerbonne</entity>, <entity>Ivan Sag</entity>, <entity>Ann Taylor</entity>, two anonymous reviewers, and*
5. *in particular <entity>Gosse Bouma</entity>, for comments, discussion and encouragement. All remaining errors are my own.*
6. *The research reported here was supported by <entity>EU TMR</entity> grant No. ERB4001GT950989.</acknowledgement>*
7. *70 Possessive affixes and complement composition PLC status is a burning one. I start therefore by considering this issue with*

Figure 1. Labeled acknowledgement passage from dataset 2, numbered to demarcate lines in the text. This particular passage contains an acknowledgement in a footnote, and the unusual structure is due to postscript to text conversion.

EXTRACTION ALGORITHM

In this section we describe the components of our acknowledgements extraction technique, including acknowledgement passage identification and entity name extraction. The algorithm uses two methods for identifying acknowledgement passages: a regular expression for extracting acknowledgement sections and a SVM for identifying lines containing acknowledgements outside of labeled acknowledgement sections. A regular expression is used to extract entity names from acknowledging text.

Passage Identification by Regular Expression

For each document, we first try to identify a labeled acknowledgement section using two distinct regular expressions. The first expression finds text before the reference (or bibliography) section or the first appendix and

after the nearest occurrence of the word "acknowledgement". If the first expression fails to find a match, the second expression finds text before the abstract or introduction and after the nearest occurrence of the word "acknowledgement". Section headers are identified by matching a newline followed by the section title or a newline followed by a number sequence (e.g. 3 or 4.5.1) followed by white space followed by the section title. Any passages over 1500 characters long are discarded at this phase of the algorithm. Although crude, this method achieves a very high degree of precision at extracting acknowledgement sections.

Passage Identification by Line Classification

The regular expression method for identifying acknowledgement sections is very precise; however, the method only retrieves roughly two thirds of the acknowledgement passages in our data set. Clearly, a more sophisticated technique for extracting the additional acknowledgement passages is needed. The following section describes a three-step technique for identifying acknowledgement passages based on line classification.

Through a careful review of 500 documents containing acknowledgement passages outside of labeled acknowledgement sections, it was found that most acknowledgements in this class of documents occur on the first page of the document or the last page preceding the reference section or the first appendix. It was further found that 87.87% of these acknowledgement passages could be found within a 100 line region in the target locations. Thus, for each document, in addition to extracting any labeled acknowledgement section, we extract the first 100 lines of the documents and the last 100 lines before the reference section or first appendix. The extracted passages are treated as candidate text for further extraction of acknowledgement passages.

We chose to use a machine learning technique for classifying lines of candidate text into the classes "acknowledging" and "non-acknowledging". A Support Vector Machine was trained to classify lines through a limited set of features, including only the unstemmed words and the number of capitalized words in the line. We used the SVM_light software package [14] for the experiment and algorithm application. The best results were obtained using a linear kernel function and all parameters set by SVM_light during cross-validation.

We obtained good classification performance for lines containing acknowledging phrases as well as capital name words. However, we found that multi-line acknowledgement passages often contain some lines that are not correctly classified by the SVM, resulting in a loss of recall during entity name extraction. This phenomenon also causes precision errors when entity names span multiple lines of different classes. In order to mitigate this problem, we merge all lines from the first line classified as positive ("acknowledging") through the last line classified as positive into a single passage, provided the positive lines are not too far apart (distance greater than 3 consecutive negatively classified lines). All lines in the passage are then reclassified as positive. This technique had the effect of improving line recall by 17.34%. In addition, line merging improved the precision of subsequent entity name extraction by 8.70%.

Entity Name Extraction

Once we obtain acknowledgement passages via the techniques discussed above, the task remains to extract the names of acknowledged entities within the passage. We treat this as a chunk identification task and use a regular expression to determine chunk boundaries. The regular expression contains six distinct cases that are applied to text chunks in order, using greedy search. That is, any text that is found to match an early case is prevented from matching subsequent cases. Table 1 presents the ordered set of cases used within our regular expression.

Table 1. Regular expression cases for entity name extraction.

Case	Description	Examples
1	Dot-delimited acronyms	D.A.R.P.A.
2	Undelimited acronyms	DARPA
3	Capital word phrases with linking words	Air Force Office of Scientific Research
4	Capital word phrases with common abbreviations	John F. Kennedy
5	Single capital words followed by organization indicator	Denny's, Inc.
6	Single capital words not at the beginning of a sentence	Microsoft

Instance merging

Oftentimes, the same organization or person may be acknowledged using multiple name variations in different acknowledgement passages. For example, the National Science Foundation may be acknowledged as "National Science Foundation", "NSF", or "N.S.F." An individual person may also be acknowledged using the individual's full name, last name only, or using the last name and any combination of other names and initials. We employ three complementary techniques for merging organization names, leaving the merging of individual person names to future work.

The first two techniques are "online"; that is, they are used during the addition of the algorithm results to our database. The most basic technique we use is to simply strip periods out of extracted names before adding the names to our database. Thus, "NSF" and "N.S.F." are treated as the same entity. In order to store acknowledgements to the full names of organizations as well as their acronyms in the same data instance slot, we allow every acknowledged entity to have a full name and an abbreviation in our database. A manually generated list of common organization name/acronym pairs is used during the database addition process. Each extracted name that is a

sequence of capital letters is compared with the list of name/acronym pairs in order to determine the expanded name.

The online name expansion process offers a partial remedy to the problem of name expansion, but it is not possible to know all organization name/acronym pairs in advance. In order to supplement the online process, an offline procedure has been developed in which the entire database is scanned for acronyms that are not associated with any expansion. The list of acronyms is then matched against full names in the database in order to automatically determine candidate expansions. Once all candidate expansions for the acronyms have been identified, the results are presented to a human user who can then decide whether to accept one of the given expansions. All positive human input results in the merging of table entries for the acronym and the identified name expansion. The new association is entered into the name/acronym pair list for use in the online name expansion process in the future.

EXPERIMENTAL RESULTS

In order to test the performance our method, we applied the algorithm components to several labeled data sets, described in Section 2. For the purpose of analysis, we divide the algorithm into two discrete methods, A and B. Each method was tested separately on different datasets and the results were combined to produce an analysis of total algorithm effectiveness.

The process of extracting acknowledgement sections via regular expression match, along with subsequent entity name extraction, is defined as Method A. This is a two-phase procedure that was applied to 600 documents containing the union of datasets 2 and 4. Method B is defined as the process of extracting acknowledgement sections from documents in which acknowledgements occur outside of labeled acknowledgement sections, along with subsequent entity name extraction. This is a three-phase procedure involving candidate passage extraction, SVM line classification, and name extraction. Method B was tested against the 600 documents comprising datasets 3 and 4. The results of each phase of Methods A and B are presented in Table 2, along with the total algorithm performance.

Table 2. Algorithm performance by component.

Procedure	Precision	Recall
Method A		
Acknowledgement passage extraction	0.9951	0.6711
Name extraction	0.7827	0.9407
Method B		
Candidate passage selection	0.2451	0.8788
SVM line classification using line merging	0.9434	0.8772
Name extraction	0.8611	0.8732
Total Algorithm	0.7845	0.8955

In order to facilitate algorithm performance analysis in terms of actual entity names, it was necessary to estimate the mean number of acknowledged entities in passages extracted by Methods A and B. It was found that the acknowledgement sections identified by Method A contained an average of 4.89 acknowledged entities, whereas acknowledgement passages identified in Method B contained an average of 2.02 acknowledged entities. Before presenting the equations for measuring our algorithm's performance it is necessary to first introduce some term definitions.

Definitions:

P_A the proportion of true acknowledgement passages that are retrieved by method A;

P_B the proportion of true acknowledgement passages that are not retrieved by method A;

N_A the average number of acknowledged entities in passages retrieved by method A, found to be 4.89;

N_B the average number of acknowledged entities in passages not retrieved by method A, found to be 2.02;

Pr_{extA} the precision of acknowledgement passage extraction in method A;

Pr_{extB} the precision of acknowledgement passage extraction in method B;

Pr_{regA} the precision of name extraction from passages retrieved in method A;

Pr_{regB} the precision of name extraction from passages retrieved in method B;

Pr_{svm} the precision of SVM line classification using line merging;

R_{extA} the recall score for method A in terms of all acknowledgement passages;

R_{regA} the recall score for name extraction from passages retrieved by method A;

R_{regB} the recall score for name extraction from passages retrieved by method B;

R_{extB} the recall score for identifying candidate text passages for classification using method B;

R_{svm} the recall score for SVM line classification using line merging.

The variables P_A and R_{extA} reference the same value, but are separated here for clarity. The equations for the total precision and recall for the complete extraction algorithm (Pr_{tot} and R_{tot}, respectively) are as follows:

$$\frac{P_A N_A}{P_A N_A + P_B N_B} Pr_{extA} Pr_{regA} + \frac{P_B N_B}{P_A N_A + P_B N_B} Pr_{svm} Pr_{regB} = Pr_{tot}$$

and

$$\frac{P_A N_A}{P_A N_A + P_B N_B} R_{extA} R_{regA} + \frac{P_B N_B}{P_A N_A + P_B N_B} R_{extB} R_{svm} R_{regB} = R_{tot}.$$

The calculations yield 0.7845 as the precision of our algorithm and a recall score of 0.8955. These numbers are

consistent with the bias toward recall coded into our regular expression.

APPLICATION TO CITESEER DATA

In this section we present the results of applying our algorithm to documents within the CiteSeer archive, as well as a prototype web interface that we created to display the acknowledgement information. We also present an analysis of the top twenty acknowledged entities within our document collection. At the time of this writing, the algorithm has been applied to 335,000 of the 575,000 documents within CiteSeer's databases.

Acknowledgement Trends

Initial analyses revealed that the distribution of acknowledgements to named entities (e.g. "National Science Foundation" or "John Smith") within the CiteSeer archive follows a power law such that only a few entities are named very frequently while a great many entities are named only rarely (see Figure 2). The power law trend in acknowledgements has been previously reported in a study involving manual extraction of acknowledgements from research papers within information science and sociology journals [7,4]. An analysis of the ISI data set [21] has shown that citations also follow a power curve. The ISI study shows an exponent of approximately -0.5 for the distribution of citations, which is comparable to our finding that CiteSeer's citation distribution follows an exponent of -0.55. Our acknowledgement data fits a power law with an exponent of -0.65, a significantly steeper slope than that exhibited by citations. We explain this by noting a high proportion of acknowledgements given to a relatively small and static list of funding agencies. These agencies fund work in many sub-communities within computer science. In contrast, we expect but have not shown that a greater number of research papers will be found within the top echelons of cited work and that citations will be shared among many classic papers according to particular scientific communities.

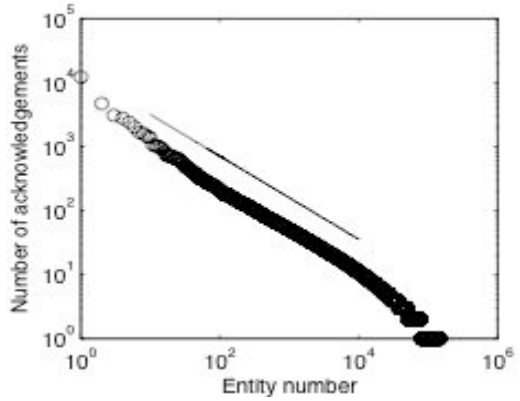


Figure 2. The distribution of acknowledgements in the CiteSeer document collection follows a power law with the exponent -0.65. A line with -0.65 slope is drawn for reference.

Top Twenty Acknowledged Entities

Table 3. Top twenty acknowledged entities in CiteSeer.

Organization	Number of acks	Total citations	C/A Metric
National Science Foundation	12,207	144,643	11.77
Defense Advanced Research Projects Agency	4,712	80,659	17.12
Office of Naval Research	3,080	48,873	15.87
Deutsche Forschungsgemeinschaft	2,780	9,782	3.52
National Aeronautics and Space Administration	2,408	21,242	8.82
Engineering and Physical Sciences Research Council	2,007	16,582	8.26
Air Force Office of Scientific Research	1,657	16,850	10.17
Natural Sciences and Engineering Research Council of Canada	1,422	12,050	8.47
International Business Machines	1,380	23,948	17.35
Department of Energy	1,054	5,562	5.28
Australian Research Council	1,010	5,464	5.41
Intel Corporation	962	14,441	15.01
Digital Equipment Corporation	831	16,390	19.72
European Union Information Technologies Program	825	9,594	11.63
Hewlett-Packard	735	11,186	15.22
National Institutes of Health	709	7,279	10.27
Army Research Office	666	7,709	11.58
Sun Microsystems	651	12, 042	18.50
Netherlands Organization for Scientific Research	646	2,843	4.40
Carnegie Mellon University	640	10,840	16.94

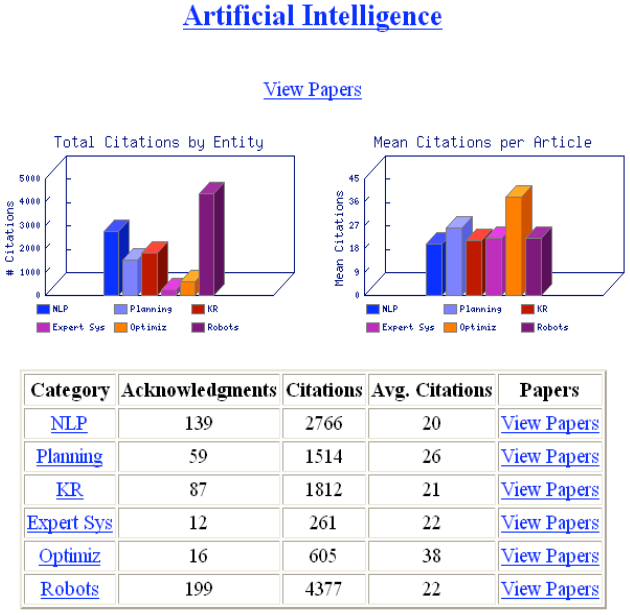
We have applied our acknowledgement extraction algorithm and name instance merging techniques to 335,000 documents within the CiteSeer archive. In this section we present the top twenty acknowledged entities within the archive, ordered by the total number of acknowledgements made to the entities. In addition to acknowledgement counts, we use the citation index in

CiteSeer to develop further measures of entity impact. Given that all but one of the top twenty entities are either science funding agencies or corporate sponsors of scientific work, we believe that the combination of acknowledgements and citation counts in order to measure relative organizational impact is particularly relevant.

In addition to total acknowledgement counts, the total number of acknowledgements made to each entity is presented along with the mean citations per acknowledging paper, which we define as the C/A Metric. We take the C/A Metric as a measure of the relative impact per acknowledgement to each entity, thus a rough measure of the effectiveness of research sponsored by the entities at disseminating into the computer science community. The results of the analysis are presented in Table 3.

Prototype Web Interface

In order to make the results of our acknowledgement extraction publicly available, we have developed a prototype web interface to our data collection. Currently, the interface supports two distinct views on the data, including an interface to the most acknowledged entities in the database as well as a detailed view of the acknowledgements to individual entities (see Figure 3 for a partial screen shot of this view). In addition, a search interface into the list of entity names is provided.



Category	Acknowledgments	Citations	Avg. Citations	Papers
NLP	139	2766	20	View Papers
Planning	59	1514	26	View Papers
KR	87	1812	21	View Papers
Expert Sys	12	261	22	View Papers
Optimiz	16	605	38	View Papers
Robots	199	4377	22	View Papers

Figure 3. A partial screen shot of the taxonomy breakdown feature of CiteSeer's prototype acknowledgements interface. This particular display pertains to acknowledgements made to DARPA within the category of artificial intelligence.

A simple taxonomy is employed to provide an initial look into more detailed analyses of entity impacts. Documents within the CiteSeer archive are classified into a shallow hierarchy of computer science topics via keyword search, based on the top-level categories of applications, architecture, artificial intelligence, hardware, compression, machine learning, human computer interaction, networking, operating systems, programming, security, software engineering, theory, databases, agents, information retrieval, and the world wide web. The taxonomy was developed by Steve Lawrence and is employed in CiteSeer's Computer Science Directory feature. The precision and recall of the taxonomy classification procedure has not been formally evaluated, so we present the taxonomy as an informal tool for providing additional organization to our data. Future versions target the inclusion of ACM taxonomic classifications as the basis for document categorization.

Each view in the interface presents a breakdown of acknowledgements in terms of the total acknowledgements in the database as well as a breakdown of acknowledgements by taxonomic category. Within the individual entity view, it is also possible to retrieve the individual acknowledging papers and browse the data via document co-acknowledgement links. The interface is planned for deployment within the next major release of CiteSeer.

SUMMARY AND FUTURE WORK

This paper describes a mixture method for automatic acknowledgement extraction using a combination of regular expressions and a SVM for acknowledgement passage identification and a regular expression for entity name extraction. The algorithm has been used to extend the semantics of research contribution within the CiteSeer Digital Library. The results obtained from the acknowledgement extraction algorithm have shown that the algorithm is a viable tool for providing acknowledgement metadata content in research libraries, and for creating initial analyses of the relative impacts of acknowledged entities in document collections. Through coupling the extraction results with the ACI capability of the CiteSeer indexing engine, we have measured the relative impacts of acknowledged entities within the CiteSeer document archive.

There are some improvements that can be made both to our algorithm and our analyses. In particular, the line merging method used in tandem with SVM line classification to extract acknowledgement passages leads in some cases to incomplete passage extraction. Since lines of acknowledging text after the last line classified as positive in an acknowledgement passage are discarded, errors of precision are induced when there are multi-line name phrases beginning on the last line of positive-classified text. Errors of recall are also induced if the names of acknowledged entities are located on discarded lines. Future work will investigate the usefulness of surrounding structural cues within PostScript or PDF files for correctly identifying the boundaries of acknowledgement passages. An incremental performance improvement could be gained for entity name extraction by developing more sophisticated name recognition software such as that described in [1,20].

Work is currently underway to improve the data that our algorithm is able to extract. By developing entity classification techniques based on machine learning tools, we plan to automatically separate entities into the categories of funding agencies, corporations, educational institutions, and individuals. In addition, we plan implement language-aware parsing methods to capture the type of acknowledgement being made, using Cronin's acknowledgement classification scheme as a base ontology. Such improvements will allow more detailed studies of acknowledgement and funding trends and facilitate the targeted analyses of specific research fields as well as individual research programs.

ACKNOWLEDGEMENTS

This work was partially supported by NSF grants 0330783 and 0202007 and by Microsoft Research. We would like to thank David Mudgett for useful discussions.

REFERENCES

[1] D.M. Bikel, S. Miller, R. Schwartz, & R. Weischedel. NYMBLE: A high-performance learning name-finder. In *Proc. of ANLP*, pp. 194-201, 1997.

[2] R.D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, 2(4). www.firstmonday.org, 1997.

[3] H.L. Chieu & H.T. Ng. A maximum entropy approach to information extraction from semi-structured and free text. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pp 786-791, 2002.

[4] B. Cronin, G. McKenzie, L. Rubio, & S. Weaver-Wozniak. Accounting for influence: Acknowledgments in contemporary sociology. *J. Am. Soc. Infom. Sci.* 44: 406-412, 1993.

[5] B. Cronin, D. Shaw, & K. La Barre. A cast of thousands: Coauthorship and subauthorship collaboration in the 20th Century as manifested in the scholarly journal literature of psychology and philosophy. *J. Am. Soc. Inf. Sci. Tec.* 54: 855-871, 2003.

[6] E. Davenport, & B. Cronin, B. Who dunnit? Metatags and hyperauthorship. *J. Am. Soc. Inf. Sci. Tec.* 52: 770, 2001.

[7] C.H. Davis, & B. Cronin. Acknowledgments and intellectual indebtedness: A bibliometric conjecture. *J. Am. Soc. Inform. Sci.* 44, 590-592, 1993.

[8] S. Dumais, J. Platt, D. Heckerman, & M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, pp 148-155, Nov., 1998.

[9] D. Edge. Quantitative measures of communication in science. *Hist. Sci.* 17: 102-134, 1979.

[10] J. Fan, K. Barker, B. Porter, & P. Clark. Representing roles and purpose. In *Proceedings of the International Conference on Knowledge Capture*, pp. 38-43, 2001.

[11] E. Garfield. Quantitative measures of communication in science. *Science* 144: 649-654, 1964.

[12] C.L. Giles & I.G. Councill. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *PNAS* 101(51): 17599-17604, 2004.

[13] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhuang, & E. Fox. Automatic document metadata extraction using Support Vector Machines. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp 37-48, May, 2003.

[14] T. Joachims. Making large-scale Support Vector Machine learning practical. In B. Scholdkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge MA, 1998.

[15] T. Joachims. A statistical learning model of text classification with Support Vector Machines. In W.B. Croft, D.J. Harper, D.H. Kraft, and J. Zobel, editors, *Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval*, pp 128-136, 2001.

[16] T. Kudoh & Y. Matsumoto. Use of support vector learning for chunk identification. In *Proc. of CoNLL-2000 and LLL-2000*, 2000.

[17] S. Lawrence, C.L. Giles, & K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67-71, 1999.

[18] K.W. McCain. Communication, competition, and secrecy: the production and dissemination research-related information in genetics. *Sci. Technol. Hum. Val.* 16: 491-516, 1991.

[19] P. McNamee & J. Mayfield. Entity extraction without language-specific resources. In D. Roth and A. van den Bosch, editors, *Proc. of CoNLL-2002*, pages 183-186, 2002.

[20] A. Mikheev, C. Groover, & M. Moens. Description of the LTG System Used for MUC-7. In *Proc. Of MUC-7*, 1998.

[21] S. Redner. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. Jour.* 4: 131-134, 1998.

[22] K. Seymore, A. McCallum, & R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction*, pp 37-42, 1999.

[23] K. Takeuchi & N. Collier. Use of Support Vector Machines in extended named entity. In D. Roth and A. van den Bosch, editors, *Proc. of the 6th Conference on Natural Language Learning*, 2004.