A System For Automatic Personalized Tracking of Scientific Literature on the Web

Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles
NEC Research Institute
Princeton, NJ 08540
{kurt,lawrence,giles}@research.nj.nec.com

ABSTRACT

We introduce a system as part of the CiteSeer digital library project for automatic tracking of scientific literature that is relevant to a user's research interests. Unlike previous systems that use simple keyword matching, CiteSeer is able to track and recommend topically relevant papers even when keyword based query profiles fail. This is made possible through the use of a heterogenous profile to represent user interests. These profiles include several representations, including content based relatedness measures. The CiteSeer tracking system is well integrated into the search and browsing facilities of CiteSeer, and provides the user with great flexibility in tuning a profile to better match his or her interests. The software for this system is available, and a sample database is online as a public service.

KEYWORDS: user profile, citation index, knowledge representation, information filtering.

INTRODUCTION

There has always been a need for humans to be kept current on important matters, but the time and effort required to do so can be enormous. Very early, this problem was handled by the creation of periodicals[1], and throughout history, the quantity and diversity of such publications has increased. In modern times, information scarcity has become information overload. In particular, the rate of publication of scientific literature grows each year, making it increasingly harder for researchers to keep up with novel relevant published work. The advent of digital libraries was a technological response to this overload. However, even with easier methods of searching through scientific documents, researchers must still expend a great deal of time and effort looking for new publications that may interest them.

[1]The first periodic newspaper is considered to be the Roman "Acta Diurna", which Julius Caesar began in about 59 B.C. [8].

Previously we have introduced *CiteSeer*, a system that performs Autonomous Citation Indexing (ACI) of scientific publications on the Web [9, 11]. CiteSeer helps users in ways that many traditional digital libraries do not. It provides the facilities to browse by citation links and allows finding both citing and cited papers of an interesting work. It summarizes citation contexts to make quick appraisal of papers easier, and it gives citation statistics including the number of citations for each cited paper and identification of self-citations. However, after spending the time to make a literature search and possibly downloading papers from the Web, the effort that the user put into the search is often forgotten and lost. Later, the user may wish to perform a search about the same topic to find new relevant papers that have appeared since the last time a search was performed. This requires a repeat of the manual labor in searching and browsing to find the papers just like the first time.

We introduce a tracking system into CiteSeer that uses profiles to represent a user's topical interests in scientific literature. CiteSeer can examine its database of publications to determine whether any new papers are related to the user's interests. If so, then the user can be alerted by e-mail or whenever they next use CiteSeer's Web based interface. CiteSeer includes, but goes beyond, simple keyword matching to determine whether a user will be interested in a new paper. A heterogeneous relatedness measure is used to identify new related documents. Also, citation links can be monitored to discover new citations to existing papers. CiteSeer not only tracks interesting papers for the user, but provides a configuration facility by which the user can change the profile to more closely reflect his or her interests.

Representing User Interests CiteSeer's tracking system acts as a proxy for user interests. It attempts to decide whether a newly available paper would be interesting enough to the user to be worth mentioning it to him or her. In order for such a system to be effective, it must be able to accurately represent a user's opinion of what is interesting. CiteSeer relies on a number of different representations of the user's notion of an interesting paper. A new paper is deemed relevant if it satisfies the requirements of any of the representations. Having a diversity of representations is important for several reasons. First, not every person searches for literature

in the same manner. Each person has their own set of techniques that they use to find useful papers. By representing a user's interests with a profile consisting of a heterogeneous set of features, a wider variety of users may be accommodated. Second, even for a single user, no one representation may be adequate to capture what the user considers interesting. A user's opinion may be complex enough that each type of representation only partially covers the user's notion of what makes a paper interesting.

CiteSeer uses two general methods for determining paper relevance: (i) constraint matching and (ii) feature relatedness. Constraint matching allows a user to describe an interesting paper by specifying constraints. This include such methods as keyword matching on the text of the paper, or metadata constraints such as specifying a source URL. Feature relatedness allows a user to specify a set of papers that are interesting, and CiteSeer tries to find papers that are related to the specified set.

CONSTRAINT MATCHING

A very simple, yet highly effective method of determining whether a paper is relevant is constraint matching. Many digital libraries allow search by features such as included keywords, age, manual classification and many other features that may be useful in determining a relevant document. CiteSeer includes several constraint matching methods that may be included in a user's profile.

Keyword Matching Although it is commonly used, keyword matching should be used with care if it is to be an effective means of detecting relevant papers. The context of the keyword is of high importance, since it has an impact on how related the keyword is to the central ideas of the paper. CiteSeer allows keyword matching, for which the context is restricted to specific parts of the document. Currently, these parts include the (i) title, (ii) header, (iii) abstract, and (iv) main body of text of the document. CiteSeer can with greater than 90% reliability determine the title of papers that it parses from Postscript (and 100% of those where the title is given, e.g. the paper is downloaded from another database). The header is everything from the beginning of the paper till the beginning of the abstract, or the first section if there is no abstract. The header often contains important information such as the author names and affiliations, and references to where the paper has been published. CiteSeer allows a user to specify keywords that will match in the header, but not in the title. For example, this will allow discrimination of the word "research" in the title versus the same word in an author's affiliation.

Keyword matching is a powerful method of identifying interesting new papers but requires the explicit determination of good keywords, which can be difficult. If poor keywords are chosen, then undesirable papers may be incorrectly identified as interesting. If good keywords are not included, then many valuable papers may be missed. Because of this, CiteSeer uses a variety of other forms of relevance in addition to keywords.

Citation Links Citation of previous published research ties a scientist's work to results from earlier research upon which it builds. To know which and how later publications cite a particular paper gives an indication of the effect of the cited work on the research community. One of the indices in a CiteSeer database is for the references at the end of the paper. Users may search for citations by keyword, and then be given links to all of the citing papers in the database, as well as the context of the citations in the main text of those papers. By presenting these citation contexts together, CiteSeer forms a summary of assessments of the cited work. If the cited work is particularly important or interesting, new citing papers may also be of interest. CiteSeer's tracking system allows a user to specify interesting citations. When new citing papers appear, the user can be informed as to their existence. One example of why a user may want to use such a representation of their interests is in citations to their own publications. Citations to a researcher's work may contain valuable feedback. Particularly for very prolific authors, keeping up with citations to one's work could be very time consuming.

Use of Metadata Since metadata is a descriptive tag associated with a document, it may provide useful information about the relevance of a scientific publication. Since CiteSeer takes most of its papers from the Web, it records the URL from which each publication is linked. Users can specify a URL to track, and when new papers appear linked from it, they are added to the list of papers to recommend. This form of relevance is important when a user wishes to keep up with publications from a particular research group or institution. The limitations of using URLs to track relevant literature mirror those of keyword matching. Extraneous URLs result in uninteresting papers being recommended and missing URLs cause useful papers to be missed. Other forms of metadata can be useful, and may be included in a future version of the CiteSeer tracking system.

RELATED PAPERS

When a user performs a literature survey on a topic of interest, he or she essentially ends up with a collection of relevant literature. By specifying keyword or metadata constraints, a user can specify an interest profile. However, there may be many other relevant publications that do not cite one of the found papers or match some other constraint. The user would like to simply say, "Tell me about new papers that are related to this one." CiteSeer is able to identify such papers through the use of relatedness measures [4].

CiteSeer attempts to capture a user's notion of relatedness between papers. This task is composed of two major challenges: (i) identifying features of the documents that represent useful semantic information, and (ii) creating functions of these features having a range space in which distances represent meaningful semantic distances. "Meaningful" in this

case is defined as "adequately represents a user's concept". We do not attempt to approach these challenges for general documents, as some information retrieval systems do. Instead, we are interested in the special case of scientific publications, which are relatively well structured, making the problem easier.

Consider a database of scientific documents created by CiteSeer. Let $\vec{F}(\cdot)$ be a set of features extractors applicable to a scientific document, and let $\vec{F}(d) = \vec{X}_d = \{x_i^d : i = 1 \ldots M_d\}$ be an M_d dimensional feature vector extracted from some document d. Let $R(\vec{X}_d, \vec{X}_e)$ be a relatedness measure between documents d and e. From the perspective of a user, we would like $R(\vec{X}_d, \vec{X}_e)$ to be small when d and e are about mostly unrelated topics and concepts, and large when d and e talk about very related issues and ideas. In this framework, challenge (i) simply amounts to choosing a good feature extractor $\vec{F}(\cdot)$ and challenge (ii) is that of choosing a useful relatedness measure $R(\cdot, \cdot)$.

Like its use of heterogeneous constraints, CiteSeer also uses a mixture of paper relatedness measures. Both text based and citation based relatedness measures are used to determine whether a paper is relevant to the user.

Text Relatedness Instead of considering a body of text to be a long string of symbols, it is common to consider a document to be a collection of words. The frequency of each unique word can be measured. A feature vector $\vec{X}_d$ is extracted from a document d where each component is one or zero to indicate the presence of a unique word or (more commonly) the frequency of the word in the document.

One often used form of this measure is known as *term frequency* $\times$ *inverse document frequency* (TFIDF) [18]. In this scheme the feature set $\vec{X}_d$ is a vector of word frequencies[2] weighted by their rarity over a collection of documents $\mathcal{D}$. Let $\mathcal{W}$ be the set of all words over $\mathcal{D}$. In a document d, let the frequency of each word stem s be f_{ds} and let the number of documents in the database having stem s be n_s. In document d let the highest term frequency be $f_{d_{max}}$. In one TFIDF scheme [17] a word weight vector element w_{ds} is calculated as:

$$w_{ds} = \frac{(0.5 + 0.5\frac{f_{ds}}{f_{d_{max}}})(\log \frac{|\mathcal{D}|}{n_s})}{\sqrt{\sum_{j \in d}((0.5 + 0.5\frac{f_{dj}}{f_{d_{max}}})^2 (\log \frac{N_\mathcal{D}}{n_j})^2)}}$$

where $|\mathcal{D}|$ is the total number of documents. Thus for TFIDF, $\vec{X}_d$ is the $|\mathcal{W}|$ dimensional vector of w_{ds} values. Once the feature vectors have been extracted for two documents, the distance between them may be calculated. Commonly, a dot product or Euclidean distance measure is used. The TFIDF relatedness between two documents d and e is a dot product of the two word vectors $\vec{X}_d$ and $\vec{X}_e$ given as:

$$R_{TFIDF}(\vec{X}_d, \vec{X}_e) = \vec{X}_d \cdot \vec{X}_e$$

CiteSeer uses the TFIDF distance between the abstracts and between text bodies of papers to determine whether a newly available paper is related to one of the papers specified by the user. If the TFIDF relatedness is above a threshold for either the abstract or full text, then it is considered relevant enough to be recommended. Currently this threshold is tuned by us by hand, but in the future could be adjusted by the user or learned from user feedback.

The total number of unique word stems $|\mathcal{W}|$ in the collection of documents $|\mathcal{D}|$ can be quite large, presenting sparsity problems for TFIDF. This has been approached in a variety of ways such as chopping off the smallest terms [17] (what CiteSeer does now) or using a dimensionality reducing mapping such as *Latent Semantic Indexing* [5].

Citation Relatedness Word based similarity measure can be useful, but do not take advantage of specific features in scientific publications. In addition to word based similarity measures, CiteSeer uses common citations to make an estimate of document relatedness. Our premise is that if two scientific papers cite some of the same previous publications, then these two papers may be related. If a cited work is very obscure then this is a more powerful indicator than if a citation is to an extremely well known and often cited publication. The measure that captures this idea of relatedness is called "Common Citation $\times$ Inverse Document Frequency" (CCIDF) [4] and is partially analogous to the word vector based TFIDF. Let c_i be the frequency of a citation i in a collection of documents $\mathcal{D}$, let $w_i = 1/c_i$ be the inverse frequency, and let $\vec{W}_\mathcal{D}$ be the vector of these inverse frequencies. Let c_{di} be a Boolean indicator of whether document d contains citation i and let $\vec{X}_d$ be the resulting Boolean vector. The CCIDF relatedness between a newly downloaded document e and a document of interest d (specified by the user) is defined as:

$$R_{CCIDF}(\vec{X}_d, \vec{X}_e) = tr(\vec{X}_d \times \vec{X}_e{}^T) \cdot \vec{W}_\mathcal{D}$$

where $tr(\cdot)$ is the trace function and $\times$ is the outer product. Any document e having a value $R_{CCIDF}(\vec{X}_d, \vec{X}_e)$ above a set threshold is considered relevant. In the future we intend to explore refinements to CCIDF to consider more information about each citation such as placement in the text body, frequency, and context.

PROFILE CREATION

A CiteSeer profile is a machine representation of a user's notion of an interesting publication. The creation of a user profile is integrated into the process of using CiteSeer's searching and browsing functions to find papers of interest. When a user uses CiteSeer through its Web browser interface, a cookie is used to assign that user a unique identifying number. This unique number allows CiteSeer to keep track of

[2] Actually, CiteSeer uses word stems generated by Porter's algorithm [15].

CiteSeer
Scientific Literature Digital Library

Home Options Edit Profile Recommendations Help Add Documents Feedback Papers About

Find: c l giles or c lee giles or c giles or l giles

Search Citations Order by: Citations Max: 100 Field: Any

Search Indexed Articles Order by: Citations Max: 10 Field: Any

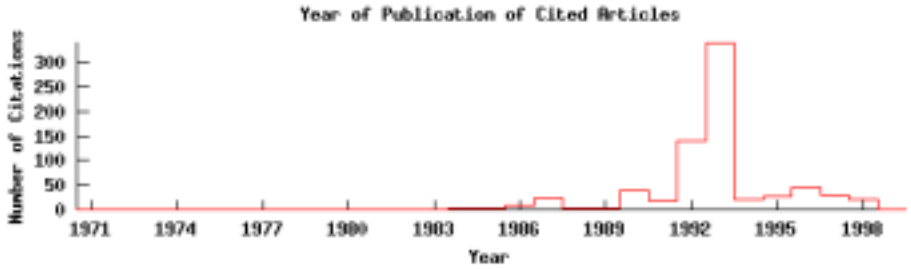
Searching for **c l giles or c lee giles or c giles or l giles** in **Computer Science** (161911 documents 2352873 citations total).

1186 citations found.

Click on the [Context] links to see the citing documents and the context of the citations. Track All Documents

Citations [hosts] (self)	Article
58 [47] (18)	C.L. **Giles**, C.B. Miller, D. Chen, H.H. Chen, G.Z. Sun, and Y.C. **Lee**. *Learning and extracting finite state automata with second--order recurrent neural networks.* Neural Computation, 4(3):393--405, 1992. Context Bib Related Track Check
29 [18] (2)	B. Hassibi and D. G. Stork. *Second-order derivatives for network pruning: Optimal brain surgeon*. In S. J. Hanson, J. D. Cowan, and C. L. **Giles**, editors, Advances in Neural Information Processing Systems, vol. 5, pages 164--171. Morgan Kauffman, San Mateo, CA, 1993. Context Bib Related Track Check
20 [16] (1)	**Giles**, C. L., Sun, G. Z., Chen, H. H., **Lee**, Y. C., and Chen, D. (1992). *Higher order recurrent networks and grammatical inference*. Neural Computation, 4(3):393--405. Context Bib Related Track Check
18 [15] (3)	P. Simard, Y. LeCun, and J. Denker. *Efficient pattern recognition using a new transformation distance*. In S. J. Hanson, J. D. Cowan, and C. L. **Giles**, editors, Advances in Neural Information Processing Systems 5, San Mateo, CA, 1993. Morgan Kaufmann. Context Bib Related Track Check
15 [10]	P. Dayan and G. E. Hinton. *Feudal reinforcement learning*. In S. J. Hanson, J. D. Cowan, and C. L. **Giles**., editors, Advances in Neural Information Processing 5, pages 271--278, San Mateo, CA, 1993. Morgan Kaufmann. Context Bib Related Track Check

[... section deleted ...]



Self-citations are not included in the graph or the main number of citations.

Figure 1: The results of a CiteSeer query for citations to “C. Lee Giles”. Different abbreviations in the query cover different ways that “C. Lee Giles” is abbreviated in citations.

the profile if the user wishes to remain anonymous. It also greatly simplifies the process of updating the profile with minimal user effort. The user may also provide an e-mail address to which recommendations may be sent periodically.

In order to actually begin building a profile, the user simply needs to specify a citation for which new citing papers will be tracked, a document for which related papers will be tracked, or a keyword or URL to add to the profile.

As a citation example, Figure 1 shows the results of a CiteSeer query for all citations to “C. Lee Giles”. In order to get a list of already existing citations to the paper “Learning and extracting finite state automata with second-order recurrent neural networks”, including contexts, the user simply chooses the **Context** link. If this citation appears interesting to the user, then the **Track New Cites** link can be chosen to add this citation to their profile. In the future, when new papers that make this citation are added to the database, they

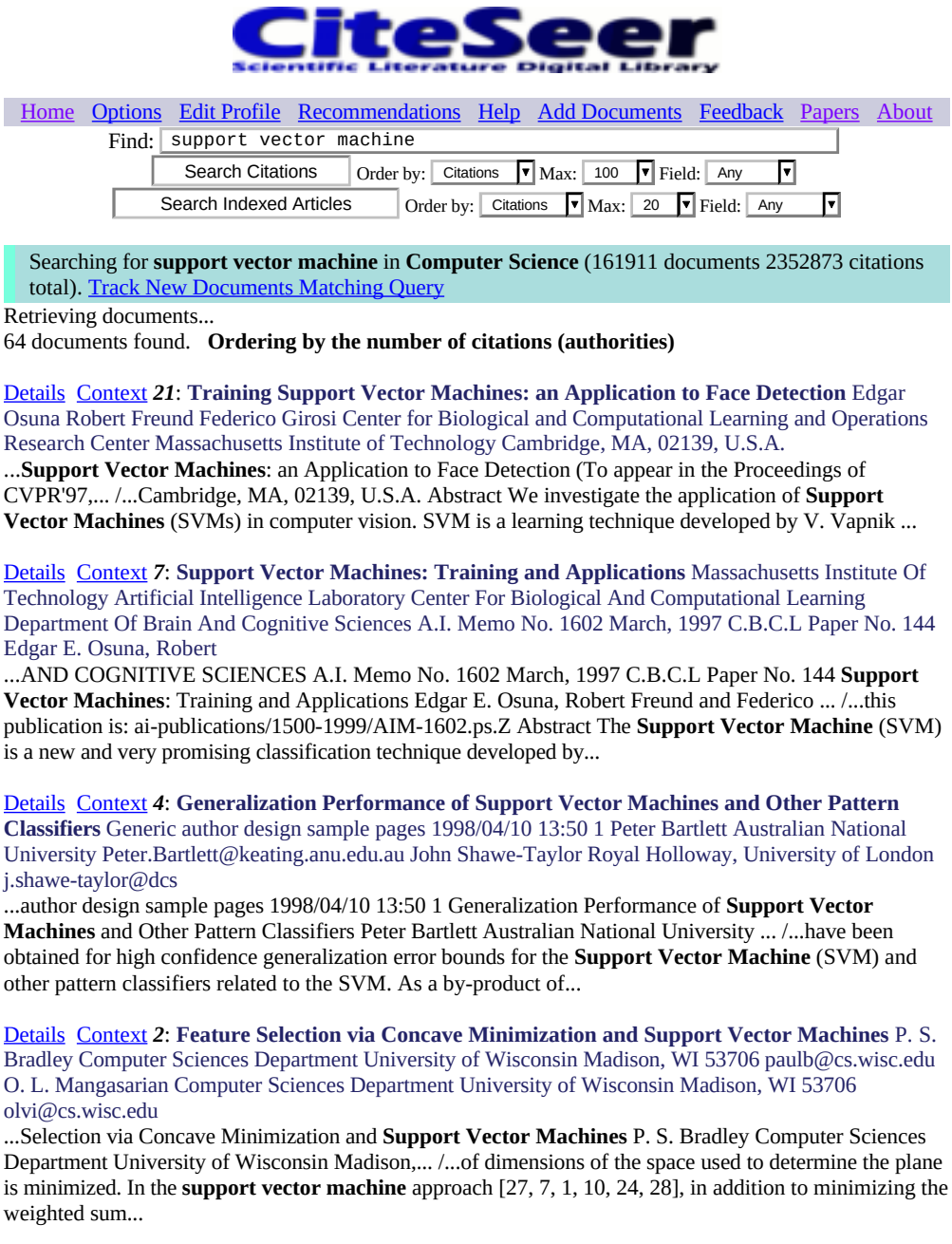


Figure 2: The results of a CiteSeer query for documents containing the term "support vector machine".

can be recommended to the user as potentially interesting.

As a document example, Figure 2 shows part of the results of a query for full documents containing the term "support vector machine". If the user is interested in the paper "Training support vector machines: An application to face detection", he or she can choose the **Details** link to get more information, as is shown in Figure 3.

The *active bibliography* section as shown in Figure 3 gives a list of documents that are related in the sense of CCIDF similarity, along with the degree of similarity. If the user wishes to track new papers that are related to this one, then the **Track Related Documents** link can be chosen to add this document to the user's profile. Additionally, the details of existing related documents can be retrieved, and new documents related to these can be tracked as well.

In order to add keywords to the profile, the user can choose the **Track New Documents Matching Query** button from the main CiteSeer search page, as shown in Figure 2. As new documents that match a given query are found, they will be recommended to the user.

Training Support Vector Machines: an Application to Face Detection

Edgar Osuna
Robert Freund
Federico Girosi
Center for Biological and Computational Learning and
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA, 02139, U.S.A.

ftp://ftp.ai.mit.edu/pub/cbcl/cvpr97-face.ps.gz Context Source HTML Track Related Documents

Abstract: We investigate the application of Support Vector **Machine**s (SVMs) in computer vision. SVM is a learning technique developed by V. Vapnik and his team (AT&T Bell Labs.) that can be seen as a new method for training polynomial, neural network, or Radial Basis Functions classifiers. The decision surfaces are found by solving a linearly constrained quadratic programming problem. This optimization problem is challenging because the quadratic form is completely dense and the memory requirements grow with the square of the number of data points. We present a decomposition algorithm that guarantees global optimality, and can be used to train SVM's over very large data sets. The main idea behind the decomposition is the iterative solution of sub-problems and the evaluation of optimality conditions which are used both to generate improved iterative values, and also establish the stopping criteria for the algorithm. We present experimental results of our implementation of SVM, and demonstrate the ...

Active bibliography (related documents):

Details Context ***0.85***: **Support Vector Machines: Training and Applications** Massachusetts Institute Of Technology Artificial Intelligence Laboratory Center For Biological And Computational Learning Department Of Brain And Cognitive Sciences A.I. Memo No. 1602 March, 1997 C.B.C.L Paper No. 144 Edgar E. Osuna, Robert

Details Context ***0.4***: **Rotation Invariant Neural Network-Based Face Detection** Henry A. Rowley Shumeet Baluja Takeo Kanade December 1997 CMU-CS-97-201 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 Justsystem Pittsburgh Research Center 4616 Henry Street Pittsburgh, PA 15213

Citations made in this document:

Context [2] G. Burel and D. Carel. *Detection and localization of faces on digital images*. Pattern Recognition Letters, 15:963--967, 1994.

Context [3] C.J.C. Burges. *Simplified support vector decision rules*. In International Conference on Machine Learning, pages 71--77. 1996.

Context [4] C. Cortes and V. Vapnik. *Support vector networks*. Machine Learning, 20:1--25, 1995.

Context [5] N. Kruger, M. Potzsch, and C. v.d. Malsburg. *Determination of face position and pose with learned representation based on labled graphs*. Technical Report 96-03, Ruhr-Universitat, January 1996.

[... section deleted ...]

Figure 3: Document details on a user selected paper.

RECOMMENDATIONS

CiteSeer uses a combination of Web search engines, Web crawling, and mailing list monitoring to continuously search for new scientific publications. As they are found, the publications are downloaded, parsed, and placed into the appropriate CiteSeer databases. There are two methods by which CiteSeer can check for new items that match the user's profile and notify the user of new recommendations. First, whenever a user begins a new session of using CiteSeer through its Web based interface, he or she can be alerted to the existence of new recommendations on the main CiteSeer page. If the user chooses to display the recommendations page, each new recommended document is displayed along with the component of their profile which was used to find that document. Figure 4 shows a demonstration of what a recommendation page looks like. In this example, the new papers that match the keywords "support vector machine" or are related to the paper "Training support vector machines: An application to face detection" are shown. If any of these recommended papers are important to the user, then they can be added to the

Home Options Edit Profile Recommendations Help Add Documents Feedback Papers About

New Document Recommendations
(for database **Computer Science**)

New documents found for the query: **support vector machine**

Details Track Related **Reducing the run-time complexity of Support Vector Machines** (To appear in ICPR'98, Brisbane, Australia, August 16-20, 1998) Edgar Osuna Federico Girosi Center for Biological and Computational Learning Massachusetts Institute of Technology Cambridge, MA 02139, USA e-mail: feosuna,girosig@ai.m

New documents related to: Edgar Osuna, Robert Freund, and Federico Girosi. *Training support vector machines: an application to face detection.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 130 -- 136, 1997.

Details Context Track Related ***0.45:*** **Face Detection with In-Plane Rotation: Early Concepts and Preliminary Results** Shumeet Baluja Justsystem Pittsburgh Research Center 4616 Henry Street Pittsburgh, PA 15213 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 baluja@jprc.com

Figure 4: New papers found by CiteSeer and recommended to the user as potentially interesting. One paper matches a keyword, while the other new paper is sufficiently related to a paper previously specified by the user as interesting.

profile. CiteSeer can also check all of the existing user profiles daily for new matches, and can e-mail those recommendations to the user if desired. If a user wishes this, then he or she needs to provide a valid e-mail address on the CiteSeer **Options** page linked off of the main page. This e-mail address can also be used to recover the user's unique ID, if their browser cookie is lost or unavailable.

PROFILE TUNING

It is easy to add new documents, citations, and keywords to a profile during the process of browsing and search in CiteSeer's Web interface. The details of the profile stay out of the way of the user, keeping CiteSeer's interface simple and elegant. However, the user may want to evaluate his or her profile and delete some of its components. By choosing the **Profile** link from the CiteSeer main page, the user is presented with their profile as in the example of Figure 5.

Shown here are the various keywords, citations, and documents that comprise the user's profile. The user can choose one or more of these components for deletion, if desired. By adding and deleting profile components, the user can tune the profile to better reflect his or her interests.

PREVIOUS WORK

The functions that CiteSeer performs fit under a lot of different research umbrellas. CiteSeer's tracking system could be thought of as an information retrieval system that performs content based information filtering. There has been a great deal of information filtering research concerning text based documents (see [6] for an overview). Important issues that have concerned researchers include document representation (e.g. [18, 5]) and classification techniques (see a comparison in [19]). The use and learning of user profiles to improve the quality of information filters has also been studied (e.g. [16, 23, 14, 3, 10]). The use of relatedness measures to find interesting publications has been cast as a routing problem, a clustering problem, and even as data mining. Much of the information filtering research has focused on the problem of general or loosely structured documents. In CiteSeer's tracking system, we take advantage of the relatively rigid structure of scientific publications to choose useful features to represent relatedness and relevance.

As far as Internet specific research, CiteSeer's tracking system is similar to the capabilities provided by several Web page tracking and location systems. Some of these systems find related Web pages using distance measures based on word vector features [13, 2, 12, 14], while others use page links to find related pages [21]. There has also been work in systems that alert users to changes in manually identified interesting Web pages (e.g. [22]).

Beyond general information filtering efforts, there are other scientific publication tracking systems. CiteSeer's tracking capabilities are partially shared by commercial tools provided by The Institute for Scientific Information (ISI) [1, 7], who provide large citation indexed databases, such as the *Science Citation Index* ®. Like CiteSeer, browsing via citation links is also possible, but citation contexts are not supported. Tracking of papers by keyword is possible through their *Discover Agent* service, but no sort of heterogenous profile information is kept, and in particular, tracking by paper

Figure 5: A sample CiteSeer profile for a user. The user can add and delete components to have it reflect the user's interests more closely.

relatedness is not supported. The e-Print archive at `http://xxx.lanl.gov/` also allows tracking of new submitted scientific literature. However, this archive requires manual classification of all submitted papers, and can only track by matching on those classifications.

CONCLUSIONS

The CiteSeer tracking system allows users to automatically keep up to date with publications on the Web that are relevant to their research interests. Users can easily define profiles consisting of a heterogenous representation of their research interests. This system finds potentially relevant papers based on these profiles and recommends them to the user via e-mail or CiteSeer's Web interface. This system is unique in its ability to find papers based on a heterogeneous measure of relatedness. The tracking interface is tightly integrated into the CiteSeer system, to minimize the effort required by users to create and tune their profile. The CiteSeer software is available at no cost for non-commercial use (contact `citeseer@research.nj.nec.com` for details), and a demonstration computer science database is available as a public service. The demonstration database can be found at `http://csindex.com/`, and indexes over 150,000 computer science articles containing over 2 million citations.

FUTURE WORK

Although the CiteSeer tracking system is a powerful means of automatically keeping up to date on research topics of interest, there are several directions in which we intend to provide enhancements. User profiles currently can only be updated manually, although this process is made very easy. We intend to explore the use of learning from implicit feedback based on use of the CiteSeer system. For example, if the value of a particular tracked document could be assessed by how many new interesting papers are found as being related to it, this information could be used to adjust sensitivity parameters for relatedness. Also, there has been a great deal of interest in collaborative filtering of text documents [20, 2]. We will investigate methods to use all of the users' profiles as a database to help enhance each individual profile. Currently,

CiteSeer only recommends new documents of interest. In the future we hope to expand this capability to the recommendation of new author names and keywords, which can then be added back into the tracking profile.

Acknowledgements

We would like to thank Shumeet Baluja, Eric Baum, Robert Cameron, Eric Glover, Haym Hirsh, Steve Hitchcock, Bob Krovetz, Andrea LaPaugh, Michael Lesk, Andrew McCallum, Michael Nelson, Craig Nevill-Manning, Brian Pinkerton, Ben Schafer, Warren Smith, and David Waltz for useful comments and suggestions regarding CiteSeer.

REFERENCES

1. Institute for Scientific Information, `http://www.isinet.com`, 1997.

2. BALABANOVIC, M. An adaptive Web page recommendation service. In *Proceedings of the First International Conference on Autonomous Agents* (February 1997).

3. BLOEDORN, E., MANI, I., AND MACMILLAN, T. R. Representational issues in machine learning of user profiles. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*. Stanford, California, March 1996.

4. BOLLACKER, K., LAWRENCE, S., AND GILES, C. L. CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the Second International Conference on Autonomous Agents* (New York, 1998), K. P. Sycara and M. Wooldridge, Eds., ACM Press, pp. 116–123.

5. DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. ., AND HARSHMAN, R. A. Indexing by latent semantic analysis. *Journal of the Society for Information Science 41* (June 1990), 391–407.

6. FALOUTSOS, C., AND OARD, D. A survey of information retrieval and filtering methods. University of Maryland Technical Report CS-TR-3514, 1995.

7. GARFIELD, E. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, New York, 1979.

8. GIFFARD, C. A. Ancient Rome's daily gazette. *Journalism History 2* (1975-1976), 106–132.

9. GILES, C. L., BOLLACKER, K., AND LAWRENCE, S. CiteSeer: An automatic citation indexing system. In *Digital Libraries 98 - The Third ACM Conference on Digital Libraries* (Pittsburgh, PA, June 23–26 1998), I. Witten, R. Akscyn, and F. M. Shipman III, Eds., ACM Press, pp. 89–98.

10. KRULWICH, B., AND BURKEY, C. Learning user information interests through extraction of semantically significant phrases. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*. Stanford, California, March 1996.

11. LAWRENCE, S., GILES, C. L., AND BOLLACKER, K. Autonomous citation indexing. *IEEE Computer 32*, 6 (1999).

12. MENCZER, F. ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery. In *Machine Learning: Proceedings of the Fourteenth International Conference* (July 1997), pp. 227–235.

13. MOUKAS, A. Amalthaea: Information discovery and filtering using a multiagent evolving ecosystem. In *Proceedings of the Conference on Practical Applications of Agents and Multiagent Technology* (April 1996).

14. PAZZANI, M., MURAMATSU, J., AND BILLSUS, D. Syskill & Webert: Identifying interesting Web sites. In *Proceedings of the National Conference on Artificial Intelligence (AAAI96)* (1996).

15. PORTER, M. F. An algorithm for suffix stripping. *Program 14* (3 1980), 130–137.

16. ROCCHIO, J. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, 1971, pp. 313–323.

17. SALTON, G., AND BUCKLEY, C. Term weighting approaches in automatic text retrieval. Tech Report 87-881, Department of Computer Science, Cornell University, 1997.

18. SALTON, G., AND YANG, C. On the specification of term values in automatic indexing. *Journal of Documentation 29* (April 1973), 351–372.

19. SCHÜTZE, H., HULL, D., AND PEDERSEN, J. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR*. 1995, pp. 229–237.

20. SHARDANAND, U., AND MAES, P. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the Conference on Human Factors in Computing Systems*. Denver, May 1995.

21. SPERTUS, E. ParaSite: Mining structural information on the Web. In *Proceeding of The Sixth International World Wide Web Conference* (April 1997).

22. STARR, B., ACKERMAN, M. S., AND PAZZANI, M. Do-I-Care: Tell me what's changed on the Web. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access* (March 1996).

23. Y. Y. YAO. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science 46* (2 1995), 133–145.