Discovering Relevant Scientific Literature on the Web

Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles, NEC Research Institute

CITESEER, AN AUTOMATIC GENERATOR OF DIGITAL LIBRARIES OF SCIENTIFIC LITERATURE, USES SOPHISTICATED ACQUISITION, PARSING, AND PRESENTATION METHODS TO ELIMINATE MOST OF THE MANUAL EFFORT OF FINDING USEFUL PUBLICATIONS ON THE WEB.

THE WEB HAS PROVED ITSELF A boon to scientific publication. It lets researchers disseminate their reports faster and at lower cost than ever before, greatly increasing the number and diversity of easily available publications. At the same time, however, the acceleration of publication has increased the perceived information overload for researchers attempting to keep abreast of relevant research in rapidly advancing fields.

Scientific literature on the Web makes up a massive, noisy, disorganized database. Unlike large, single-source databases such as a corporate customer database, the Web database draws from many sources, each with its own organization. Also, owing to its diversity, most records in this database are irrelevant to an individual researcher. Furthermore, the database is constantly growing in content and changing in organization. All these characteristics make the Web a difficult domain for knowledge discovery.

To quickly and easily gather useful knowledge from such a database, users need the help of an information-filtering system that automatically extracts only relevant records as they appear in a stream of incoming records.[1] To this end, we have developed the CiteSeer digital library system.[2] CiteSeer, a custom-digital-library generator, performs information-filtering and knowledge-discovery functions that keep users up-to-date on relevant research. CiteSeer uses a three-stage process: database creation and feature extraction, personalized filtering of new publications, and personalized adaptation and discovery of interesting research and trends. These functions are interdependent—information filtering affects what is discovered, and useful discoveries tune the information filtering.

Database creation and feature extraction

The body of scientific literature on the Web spreads over many sites. It is usually in formats such as PostScript or PDF that are typically not indexed by Web search engines, and it is organized differently at each site. CiteSeer's first stage extracts features from this source to build a digital library and provides useful tools for finding literature in this library. This stage uses several heuristics that tune the process to the internal organization of scientific literature, and it sets the stage for more sophisticated adaptive filtering and discovery.

CiteSeer creates a database by downloading Web publications in a general research area—for example, neural networks or computer vision. After downloading a document, CiteSeer extracts the raw text and parses it to find fields common to most research papers: title, abstract, word frequencies, and citation list. Then it indexes these features and places them in a local database.

Instead of simple template matching, CiteSeer uses sophisticated algorithms to parse a wide variety of research paper formats. For example, a reliable method for identifying a paper's title involves finding the largest font on the first page. Also, citations to one paper

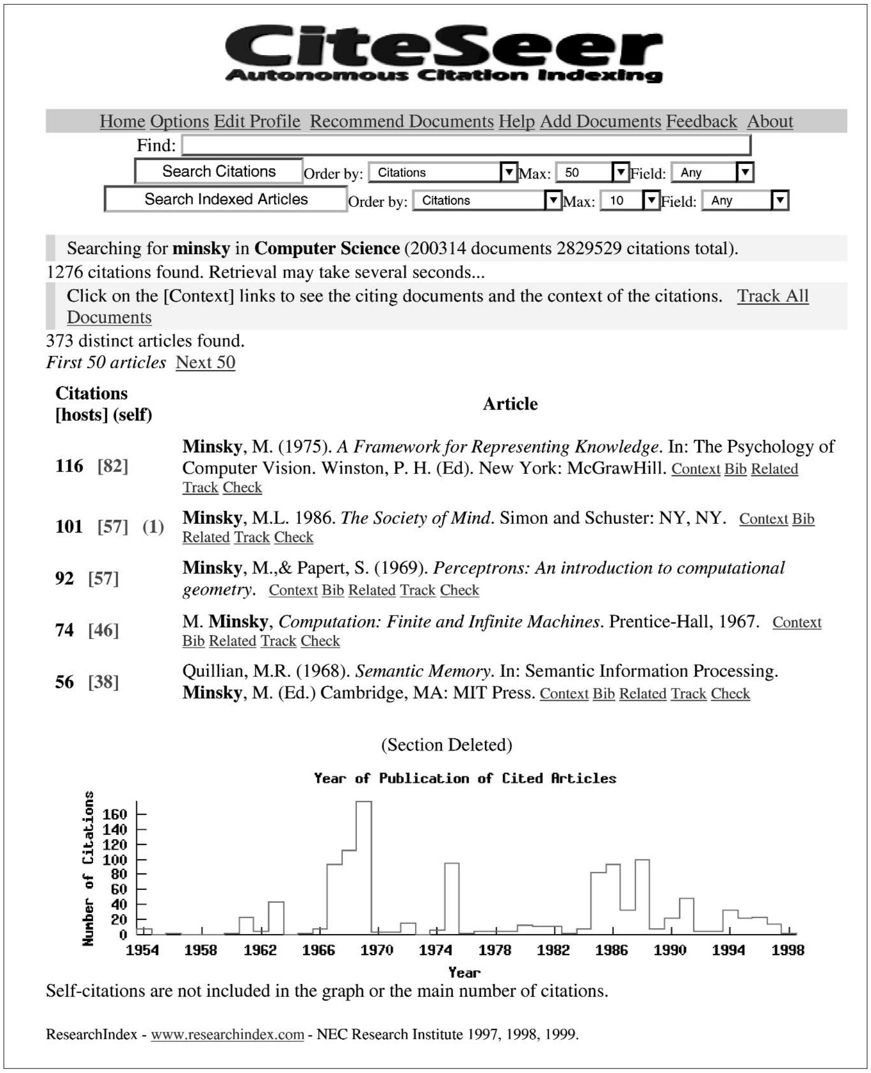
1094-7167/00/$10.00 © 2000 IEEE

might be in different formats, depending on the citing paper, so CiteSeer uses algorithms that reliably identify them as the same citation.[3] Because both a paper and citations to that paper might be in the database, title and author matching and other heuristics can automatically tie a paper to its citations. This allows CiteSeer to build a full graph of citing and cited papers.

CiteSeer provides a variety of static searching and browsing capabilities that greatly reduce the effort required to perform a literature survey.[4] Beyond traditional keyword search of the text and citations, CiteSeer provides facilities for browsing forward and backward through citation links, letting the user find both papers that cite a given article, and papers that a given article cites. CiteSeer extracts and summarizes citation contexts to facilitate quick appraisal of papers, identifies self-citations, and gives statistics including the number of citations for each paper.

Users can perform searches on downloaded documents using CiteSeer's browser-based interface. For example, Figure 1 illustrates a search for citations of an author named Minsky. A CiteSeer user performed this query on a small database of computer science papers (approximately 200,000 documents containing 2.8 million citations).

As another example, suppose the user wishes to find papers about support vector machines in the same database. CiteSeer responds to the query "support vector machine" with a list of papers ranked by number of citations, as Figure 2a shows. A user interested in the paper "Training Support Vector Machines: an Application to Face Detection" can choose the Details link to get more information. Figure 2b shows the first part of these details.


CiteSeer
Autonomous Citation Indexing
Home Options Edit Profile Recommend Documents Help Add Documents Feedback About
Search Citations
Search Indexed Articles

Find:

Order by: Citations Max: 50 Field: Any

Order by: Citations Max: 10 Field: Any

Searching for **minsky** in **Computer Science** (200314 documents 2829529 citations total).

1276 citations found. Retrieval may take several seconds...

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

373 distinct articles found.

First 50 articles Next 50

Citations **[hosts] (self)**	**Article**
116 [82]	**Minsky**, M. (1975). *A Framework for Representing Knowledge*. In: The Psychology of Computer Vision. Winston, P. H. (Ed). New York: McGrawHill. Context Bib Related Track Check
101 [57] (1)	**Minsky**, M.L. 1986. *The Society of Mind*. Simon and Schuster: NY, NY. Context Bib Related Track Check
92 [57]	**Minsky**, M.,& Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Context Bib Related Track Check
74 [46]	M. **Minsky**, *Computation: Finite and Infinite Machines*. Prentice-Hall, 1967. Context Bib Related Track Check
56 [38]	Quillian, M.R. (1968). *Semantic Memory*. In: Semantic Information Processing. **Minsky**, M. (Ed.) Cambridge, MA: MIT Press. Context Bib Related Track Check

(Section Deleted)

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

ResearchIndex - www.researchindex.com - NEC Research Institute 1997, 1998, 1999.

Figure 1. Results of a CiteSeer query for citations of "Minsky."

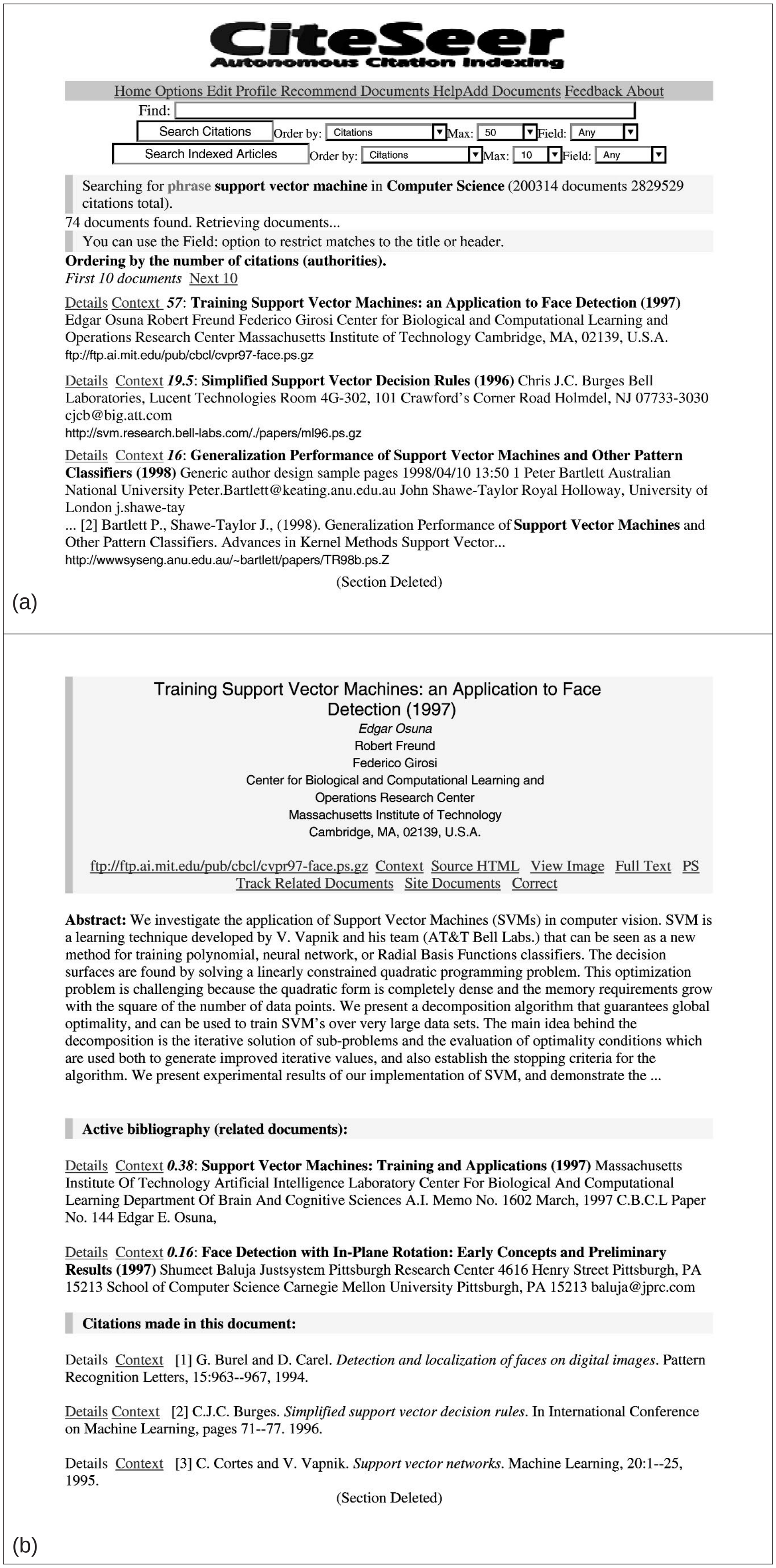
Personalized filtering

CiteSeer uses a personal profile representing a user's research interests to track and recommend relevant research. CiteSeer examines the local publication database to find new papers that might be interesting to the user and alerts the user by e-mail or through a Web-based interface. The profile adapts to the user's research interests through a feedback system using manual profile adjustment and machine learning. To modify the profile, CiteSeer watches the user's browsing behavior and the user's responses to its recommendations. These modifications might result in new recommendations, to which the user again responds. Over time, this learning cycle enables CiteSeer to find relevant papers more accurately and reliably.

Profile creation. While using CiteSeer's Web interface, users contribute to their profiles either explicitly by manually editing the profile or implicitly by browsing the database. Either action creates or modifies profile components we call pseudodocuments, which represent users' research interests. Pseudodocuments are placeholders for a set of values representing features (often only a single feature or a few features) extracted from publications. Which features to extract to form a pseudodocument is an active research area.[5,6]

CiteSeer uses a heterogeneous set of pseudodocuments including features such as keywords, URLs, citations, word vectors, and citation vectors. Evidence suggests that this set is more powerful than any single representation.[7,8] For example, research shows that retrieval based on citations often has little overlap with retrieval based on keywords.[9] Thus, a user's profile consists of a set $\mathcal{D}$ of different types of pseudodocuments. In addition to a feature value, each pseudodocument d has a weight w_d corresponding to its influence. For example, high positive w_d values mean the pseudodocument is a very good example of the user's interest, and a negative value indicates an item the user would avoid.

Figure 3 shows CiteSeer's user facility for explicitly creating a profile. From this Web page, a user can add or modify the influence of keyword or URL feature values for constraint matching. The user can also modify the influence of citations or papers previously specified while browsing. For the example profile shown here, the user selected the Track Related Documents link in Figure 2b.



Searching for phrase **support vector machine** in **Computer Science** (200314 documents 2829529 citations total).

74 documents found. Retrieving documents...

You can use the Field: option to restrict matches to the title or header.

Ordering by the number of citations (authorities).

First 10 documents Next 10

Details Context ***57***: **Training Support Vector Machines: an Application to Face Detection (1997)** 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

Details Context ***19.5***: **Simplified Support Vector Decision Rules (1996)** Chris J.C. Burges Bell Laboratories, Lucent Technologies Room 4G-302, 101 Crawford's Corner Road Holmdel, NJ 07733-3030 cjcb@big.att.com
http://svm.research.bell-labs.com/./papers/ml96.ps.gz

Details Context ***16***: **Generalization Performance of Support Vector Machines and Other Pattern Classifiers (1998)** Generic author design sample pages 1998/04/10 13:50 1 Peter Bartlett Australian National University Peter.Bartlett@keating.anu.edu.au John Shawe-Taylor Royal Holloway, University of London j.shawe-tay
... [2] Bartlett P., Shawe-Taylor J., (1998). Generalization Performance of **Support Vector Machines** and Other Pattern Classifiers. Advances in Kernel Methods Support Vector...
http://wwwsyseng.anu.edu.au/~bartlett/papers/TR98b.ps.Z

(Section Deleted)

Training Support Vector Machines: an Application to Face Detection (1997)

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 View Image Full Text PS
Track Related Documents Site Documents Correct

Abstract: We investigate the application of Support Vector Machines (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 ***0.38***: **Support Vector Machines: Training and Applications (1997)** 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,

Details Context ***0.16***: **Face Detection with In-Plane Rotation: Early Concepts and Preliminary Results (1997)** 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

Citations made in this document:

Details Context [1] G. Burel and D. Carel. *Detection and localization of faces on digital images*. Pattern Recognition Letters, 15:963--967, 1994.

Details Context [2] C.J.C. Burges. *Simplified support vector decision rules*. In International Conference on Machine Learning, pages 71--77. 1996.

Details Context [3] C. Cortes and V. Vapnik. *Support vector networks*. Machine Learning, 20:1--25, 1995.

(Section Deleted)

Figure 2. (a) The first few results of a CiteSeer query for documents containing the term "support vector machine"; (b) the top part of the document details for the first paper listed in Figure 2a.

The interestingness of new papers. CiteSeer treats a new paper d^* in the database as a pseudodocument with features corresponding to the union of the feature types in the user's profile $\mathcal{D}$. CiteSeer compares this pseudodocument with those in the profile to find a level of similarity $I_{\mathcal{D}}(d^*)$, which represents the paper's interestingness or relevance to the user. We calculate interestingness as the weighted sum

$$I_{\mathcal{D}}(d^*) = \sum_{d \in \mathcal{D}} w_d R_d(d, d^*)$$

where $R_d(d, d^*)$ is the similarity or relatedness between pseudodocument d in the user's profile and the new paper pseudodocument d^*. We weight each relatedness measure by the profile pseudodocument's influence. CiteSeer recommends new papers with $I_{\mathcal{D}}(d^*)$ greater than a certain threshold. Currently, this threshold is set at a small positive number, but we plan to allow user adjustments, as described later.

The relatedness measure $R_d(d, d^*)$ depends on the type of pseudodocuments being compared. For example, the user can create pseudodocuments as explicitly specified keywords, citations, and other constraint values. For a constraint, the appropriate relatedness measure is a simple zero or one, depending on whether the new paper matches the constraint. Although constraint-based similarity is useful, a user often wants to find papers that are related even if they do not match a given constraint. In other words, the user would like simply to say, "Tell me about new papers related to these existing papers."

A measure that captures this idea of relatedness is *common citation × inverse document frequency*—the sum of the inverse frequencies of the common citations between two papers.[4] CCIDF is similar to bibliographic coupling and is partly analogous to the word-vector-based measure called *term frequency × inverse document frequency* (TFIDF).[10] See the sidebar for more details on these relatedness measures.

Recommending papers. Once CiteSeer creates a profile, it periodically or on demand checks its database for new papers it should recommend to the user. It sends such recommendations by e-mail if the user so desires and presents them when the user chooses the Recommend Documents link, as Figure 4 shows. The recommendation ranks papers by their $I_{\mathcal{D}}(d^*)$ values and includes an explanation for each recommendation. In Figure 4, the explanation is that the recommended document is related to a paper in the profile: “Training Support Vector Machines: an Application to Face Detection.” The user can view, download, ignore, or add any recommended paper to the user profile.

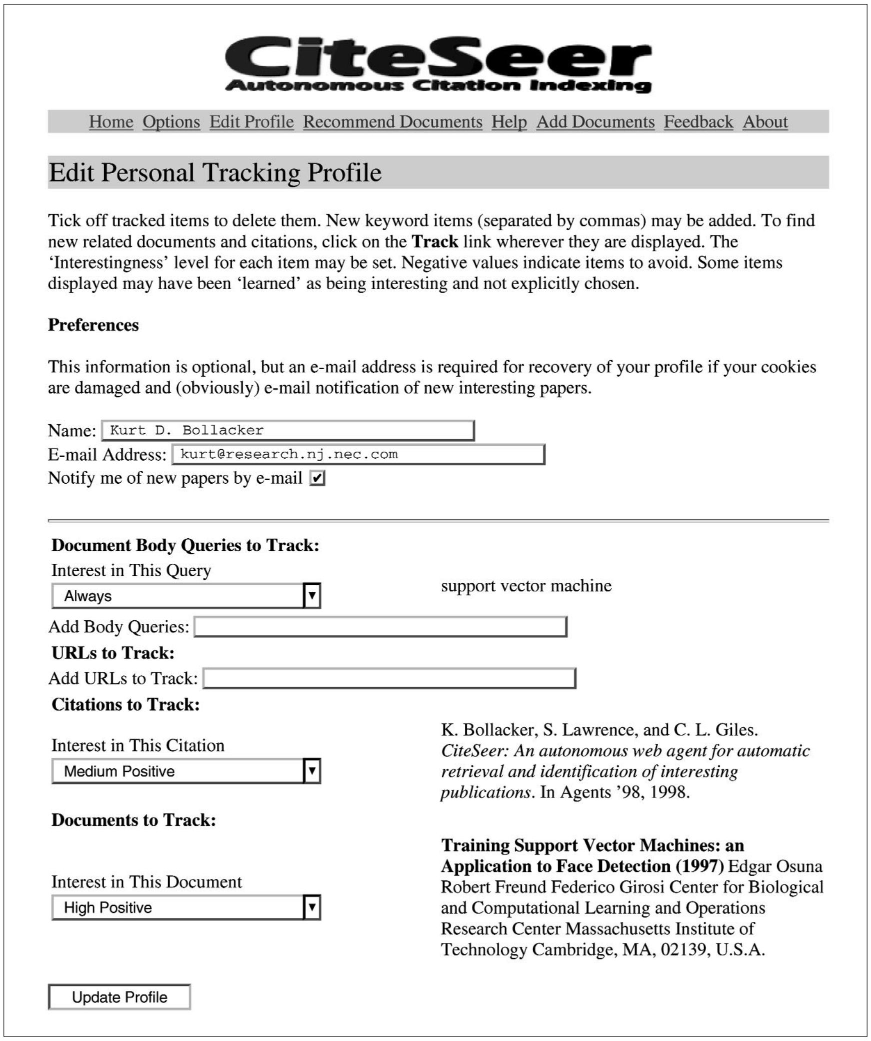
Profile adaptivity. CiteSeer adapts profiles to better represent users’ interests by modifying the pseudodocument weights w_d. It does this in three ways: observing user behavior during database browsing, allowing manual adjustment, and learning from user responses to recommendations. CiteSeer can use several types of user actions as implicit indications of interest.[11] These include viewing details, downloading a paper, and explicitly adding or removing a paper to or from the profile. For example, by viewing a paper’s details (as in Figure 2b), the user adds a CCIDF pseudodocument for that paper to the profile. Each user action b_d on pseudodocument d initializes or adds to that document’s influence (w_d) an amount corresponding to the interestingness $a(b_d)$ indicated by that action. Table 1 lists the relative $a(b_d)$ values of the various types of actions. We set these values in an ad hoc manner, and they are fixed in the current CiteSeer implementation. We consider the special case of explicitly adding to or modifying pseudodocuments to be a manual adjustment of the profile. However, manual adjustments of the w_d values are also allowed.

After recommending document d^*, CiteSeer observes the user’s response and updates the weight for each pseudodocument d in profile $\mathcal{D}$ accordingly. The update rule is

$$w_d \Leftarrow w_d + \eta a(b_{d^*})\, R_d(d, d^*)$$

where η is a learning rate and $R_d(d, d^*)$ is the relatedness measure for the specific type of pseudodocument d. This simple update rule has several useful properties:

- Weights on pseudodocuments that contribute to good recommendations increase, and weights on pseudodocuments that contribute to bad recommendations decrease.
- The system’s overall precision and recall threshold adapt to user needs implicitly and automatically. If the threshold is too low, CiteSeer recommends too many irrelevant documents, which the user ignores, thus lowering the w_d values and in turn raising the threshold. If the threshold is too high, the system recommends too few documents, thus encouraging the user to add more pseudodocuments.


CiteSeer
Autonomous Citation Indexing
Home Options Edit Profile Recommend Documents Help Add Documents Feedback About
Update Profile

Edit Personal Tracking Profile

Tick off tracked items to delete them. New keyword items (separated by commas) may be added. To find new related documents and citations, click on the **Track** link wherever they are displayed. The ‘Interestingness’ level for each item may be set. Negative values indicate items to avoid. Some items displayed may have been ‘learned’ as being interesting and not explicitly chosen.

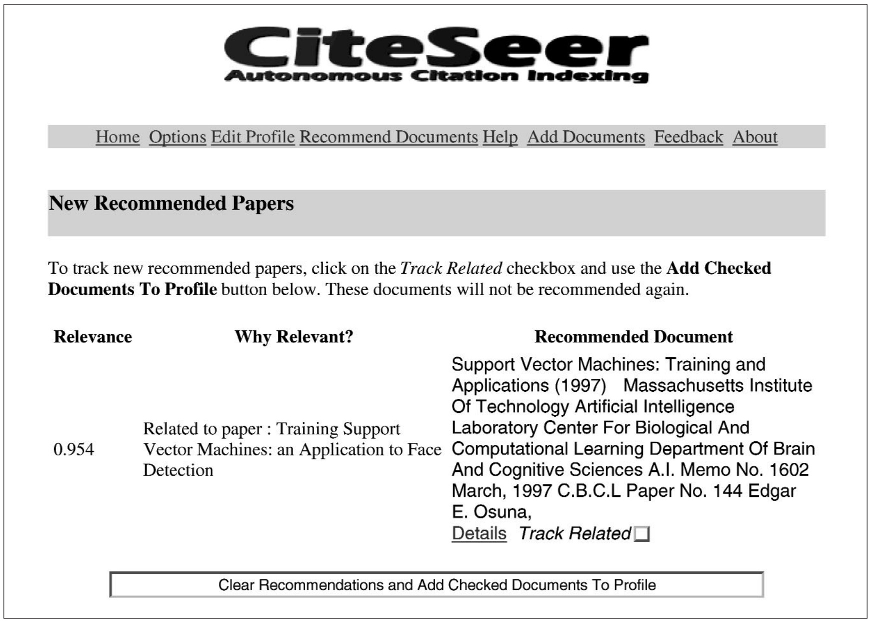
Preferences

This information is optional, but an e-mail address is required for recovery of your profile if your cookies are damaged and (obviously) e-mail notification of new interesting papers.

Name: Kurt D. Bollacker
E-mail Address: kurt@research.nj.nec.com
Notify me of new papers by e-mail ☑

Document Body Queries to Track:
Interest in This Query: Always ▾ — support vector machine
Add Body Queries:
URLs to Track:
Add URLs to Track:
Citations to Track:
Interest in This Citation: Medium Positive ▾ — K. Bollacker, S. Lawrence, and C. L. Giles. *CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications*. In Agents ’98, 1998.
Documents to Track:
Interest in This Document: High Positive ▾ — **Training Support Vector Machines: an Application to Face Detection (1997)** 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.

Figure 3. A CiteSeer user profile. Users can create components and manually adjust the influence of components to reflect their interests.


CiteSeer
Autonomous Citation Indexing
Home Options Edit Profile Recommend Documents Help Add Documents Feedback About
Clear Recommendations and Add Checked Documents To Profile

New Recommended Papers

To track new recommended papers, click on the *Track Related* checkbox and use the **Add Checked Documents To Profile** button below. These documents will not be recommended again.

Relevance	Why Relevant?	Recommended Document
0.954	Related to paper : Training Support Vector Machines: an Application to Face Detection	Support Vector Machines: Training and Applications (1997) 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, Details *Track Related* □

Figure 4. A new-paper recommendation. Recommendations include the paper’s $I_{\mathcal{D}}(d^*)$ value—in this case, 0.954—and an explanation of why the paper is recommended.

- The update rule weights the influence of different relatedness measures separately. As a result, CiteSeer can use documents in the profile that are interesting in only some ways—for example, their citation lists—to find good candidate documents using only those ways. Relatedness measures that correlate poorly with $a(b_d)$ will tend to have little influence.
- The update rule uses both explicit and implicit feedback from the user. Explicit feedback is much easier to use and more accurate, although much harder to acquire than implicit feedback.
- The model is computationally scalable. The costs of interestingness calculations and profile updates are linear with the profile's size and do not increase with the database's size.
- Developers can easily add new relatedness measures and corresponding pseudodocument types to the system.

Personalized knowledge discovery

Potentially, CiteSeer's profile adaptivity through manual adjustment and machine learning can provide more than a way to find and recommend better papers. Once a profile is well tuned to a user's interests, knowledge discovery techniques should make it possible to find new research concepts and trends that might interest the user.

New concepts. CiteSeer increases the weights of pseudodocuments that contribute greatly or often to good recommendations. Correlations between these highly weighted pseudodocument values and other feature values extracted from the same papers might reveal interesting new concepts. For example, CiteSeer could suggest author names that correlate highly with citations made by papers in the user's profile but are not already part of a constraint-based pseudodocument. If the user agrees with the suggestion, this new knowledge could be added to the user's profile to improve future recommendations.

Changes with time. Over time, a user's interests might change and grow, requiring more frequent and more substantial updates of the profile than its initial tuning to a specific interest. Papers the user added to the profile from a new research area might be unrelated to existing papers in the profile. This would tend to result in multiple interest clusters, which traditional clustering techniques should be able to discover.

New research areas. The user's profile might not contain authors or keywords with which CiteSeer can discover papers from a potentially interesting new research area. These papers, however, probably cite previously published research. If some of these citations refer to papers in the user's profile or show sufficient relatedness to papers in the profile, CiteSeer can recommend the new papers. Thus, citation-based features can be instrumental in discovering new research trends.

ALTHOUGH CITESEER HAS ALready informally proved itself very useful (the demonstration system has served millions of requests), we have much work to do. We plan to formally evaluate how well the user profiles learn and represent changes in user interests. This evaluation will include techniques such as cross-validation using random partitioning of the profile into training and test sets of pseudodocuments.

To provide better identification of personally important research trends, we intend to explore more sophisticated analysis and knowledge discovery techniques. For example, we might treat a CiteSeer database as a directed graph in which citations are edges and papers are nodes. Citation graph analysis could result in better relatedness measures or in the discovery of structural features such as citation cliques by mapping to an author citation graph. Also, technologies such as collaborative filtering might increase CiteSeer's ability to find interesting papers that would otherwise be missed.

A demonstration CiteSeer database of more than 250,000 computer science research papers containing more than three million citations is publicly available at csindex.com. We encourage you to use this free service and provide us feedback.

Table 1. A paper's interestingness as determined by user actions on that paper.

User action b_d	Document interestingness $a(b_d)$
Explicitly added to profile	Very high positive
Downloaded	High positive
Viewed details	Moderate positive
Ignored	Low negative
Removed from profile	Set to zero

Acknowledgments

We thank Eric Glover, Gary Flake, and Nelson Amaral for their helpful suggestions and comments.

References

1. C. Faloutsos and D. Oard, *A Survey of Information Retrieval and Filtering Methods*, Tech. Report CS-TR-3514, Computer Science Dept., Univ. of Maryland, College Park, Md., 1995.
2. S. Lawrence, C. Lee Giles, K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *Computer*, Vol. 32, No. 6, June 1999, pp. 67–71.
3. S. Lawrence, K. Bollacker, and C.L. Giles, "Autonomous Citation Matching," *Proc. Third Int'l Conf. Autonomous Agents*, ACM Press, New York, 1999, p. 392.
4. K. Bollacker, S. Lawrence, and C.L. Giles, "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications," *Proc. Second Int'l Conf. Autonomous Agents*, ACM Press, New York, 1998, pp. 116–123.
5. E. Bloedorn, I. Mani, and T.R. MacMillan, "Representational Issues in Machine Learning of User Profiles," *Proc. AAAI Spring Symp. Machine Learning in Information Access*, Amer. Assoc. for Artificial Intelligence, Menlo Park, Calif., 1996.
6. B. Krulwich and C. Burkey, "Learning User Information Interests through Extraction of Semantically Significant Phrases," *Proc. AAAI Spring Symp. Machine Learning in Information Access*, Amer. Assoc. for Artificial Intelligence, Menlo Park, Calif., 1996.
7. M. Balabanovic, "An Adaptive Web Page Recommendation Service," *Proc. First Int'l Conf. Autonomous Agents*, ACM Press, New York, 1997, pp. 378–385.
8. B.T. Bartell, G.W. Cottrell, and R.K. Belew, "Automatic Combination of Multiple Ranked

Retrieval Systems,” *Proc. 17th Ann. Int’l ACM-SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, New York, 1994, pp. 173–181.

9. K. McCain, “Descriptor and Citation Retrieval in the Medical Behavioral Sciences Literature: Retrieval Overlaps and Novelty Distribution,” *J. Amer. Soc. of Information Science*, Vol. 40, No. 2, Mar. 1989, pp. 110–114.

10. G. Salton and C. Yang, “On the Specification of Term Values in Automatic Indexing,” *J. Documentation*, Vol. 29, No. 4, Apr. 1973, pp. 351–372.

11. D.M. Nichols, “Implicit Rating and Filtering,” *Proc. Fifth DELOS Workshop on Filtering and Collaborative Filtering*, European Consortium for Informatics and Mathematics, Sophia Antipolis, France, 1997, pp. 31–36.

Kurt D. Bollacker is the technical director of the Internet Archive. His research interests are artificial intelligence, user profile modeling, digital libraries, and automatic knowledge discovery from large data sets. Previously, he worked at NEC Research Institute, where he participated in the CiteSeer project. He received a PhD in computer engineering from the University of Texas at Austin, where he was a National Science Foundation Graduate Fellow. He is a member of the IEEE, the ACM, and the AAAI. Contact him at Internet Archive, PO Box 29244, San Francisco, CA 94129; kurt@archive.org.

Steve Lawrence is a research scientist in computer science at NEC Research Institute. His research interests include information retrieval, digital libraries, and machine learning. His awards include an NEC Research Institute excellence award, a Queensland University of Technology medal and award for excellence, and three successive prizes in the annual Australian Mathematics Competition. He received a BSc and a BEng from Queensland University of Technology, Australia, and a PhD from the University of Queensland, Australia. Contact him at NEC Research Inst., 4 Independence Way, Princeton, NJ 08540; lawrence@research.nj.nec.com.

C. Lee Giles is a senior research scientist in computer science at NEC Research Institute. He is also an adjunct faculty member at the Institute for Advanced Computer Studies at the University of Maryland and an adjunct professor of computer and information science at the University of Pennsylvania. His research interests include intelligent information retrieval and processing, new machine-learning and AI applications in Web computing, and fundamental intelligent-system models. He is an IEEE Fellow and a member of the AAAI, ACM, INNS, OSA, AAAS, and Center for Discrete Mathematics and Theoretical Computer Science at Rutgers University. Contact him at NEC Research Inst., 4 Independence Way, Princeton, NJ 08540; giles@research.nj.nec.com.

Research paper relatedness measures

When a new candidate paper appears, CiteSeer must decide whether to recommend it to the user. If the paper is sufficiently similar to the collection of pseudodocuments making up a user’s profile, CiteSeer considers it related and recommends it. Generally, $R_d(d, d^*)$ measures relatedness between pseudodocument d and candidate document d^*. Each of the following relatedness measures is specific to the type of pseudodocument for which CiteSeer uses it.

Constraint-based relatedness. CiteSeer uses constraint-based relatedness with pseudodocuments in the profile that are not part of a paper. For example, a user specifies the term “support vector machine” as a keyword. The pseudodocument d that represents this specification is an artificial document that has this keyword as its only feature. If a candidate document d^* contains this keyword, $R_d(d, d^*)$ is unity; otherwise, it is zero.

TFIDF: word vector relatedness. Automatic retrieval systems commonly treat a document as a collection of words about which we can gather statistics. For example, we can measure the frequency of each unique word stem. (Word stemming attempts to match words by removing common endings—for example, removing “ing” and “ed” from “publishing” and “published.”) We extract a feature vector $\mathbf{W}_D$ and use it as a pseudodocument d in which each component is the frequency of a word stem in the document.

An often-used form of this measure is *term frequency × inverse document frequency.*[1] In this scheme, the feature set $\mathbf{W}_D$ is a vector of word frequencies weighted by their rarity over a document collection. Let’s say that $\mathcal{W}$ is the set of all unique words in the CiteSeer database. In pseudodocument d, let the frequency of each word stem s be f_{ds}, and let the number of documents in the database containing stem s be n_s. In document d, let the highest term frequency be $f_{d\max}$. One TFIDF scheme[2] calculates a word weight vector element w_{ds} as

$$w_{ds} = \frac{\left(0.5+0.5\frac{f_{ds}}{f_{d_{\max}}}\right)\left(\log\frac{N}{n_s}\right)}{\sqrt{\sum_{j\in d}\left(\left(0.5+0.5\frac{f_{dj}}{f_{d_{\max}}}\right)^2\left(\log\frac{N}{n_j}\right)^2\right)}}$$

where N is the total number of documents. For TFIDF, the relatedness measure based on the $|\mathcal{W}|$-dimensional vector of w_{ds} values is

$$R_d(d, d^*) = \mathbf{W}_d \cdot \mathbf{W}_{d^*}$$

Citation-based relatedness. CiteSeer uses common citations to estimate document relatedness. Our premise is that if two scientific papers cite some of the same previous publications, the two papers might be related. A very obscure cited work is a more powerful indicator than a citation to a well-known and often-cited publication. The measure we call *common citation × inverse document frequency* measures this kind of relatedness.[3] Let’s say that f_i is the frequency of a citation i in the CiteSeer database, $C_i = 1/f_i$ is the inverse frequency, and $\boldsymbol{C}$ is the vector of these inverse frequencies. We let c_{di} be a Boolean indicator of whether pseudodocument d contains i, and $\mathbf{X}_d$ be the resulting Boolean vector. We define the CCIDF relatedness of a candidate pseudodocument d^* and pseudodocument d in the profile as

$$R_d\,(d, d^*) = tr(\mathbf{X}_d \times \mathbf{X}_{d^*}) \cdot C$$

where $tr()$ is the trace function and × is the outer product.

References

1. G. Salton and C. Yang, “On the Specification of Term Values in Automatic Indexing,” *J. Documentation*, Vol. 29, No. 4, Apr. 1973, pp. 351–372.

2. G. Salton and C. Buckley, *Term Weighting Approaches in Automatic Text Retrieval*, Tech. Report 87-881, Dept. of Computer Science, Cornell Univ., Ithaca, N.Y., 1997.

3. K. Bollacker, S. Lawrence, and C.L. Giles, “CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications,” *Proc. Second Int’l Conf. Autonomous Agents*, ACM Press, New York, 1998, pp. 116–123.