Kurt D. Bollacker, Steve Lawrence, C. Lee Giles. *Discovering Relevant Scientific Literature on The Web*, IEEE Intelligent Systems, Volume 15, Number 2, pp. 42–47, 2000.

Discovering Relevant 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

Due to the ease of electronic dissemination, the world of scientific literature on the Web has grown rapidly, becoming a large, highly current database of published research. This acceleration of publication has exacerbated the difficulty researchers face keeping up to date on relevant new research trends. We believe that automatic tools to help researchers keep up with the latest relevant publications will be increasingly important in the future. One such tool, *CiteSeer*, is an automatic generator of scientific literature databases. CiteSeer uses sophisticated acquisition, parsing, and presentation methods to eliminate most of the manual effort required to perform a literature survey of publications on the Web. It also includes a personalized recommendation system that uses browsing behavior and automatic learning to adapt to individual research interests, even as they change over time. CiteSeer can pro-actively recommend new relevant research papers as they appear on the Web as well as discover new citations, keywords, and authors that may be indicative of novel research trends of interest to the user.

1 Introduction

The World Wide Web (Web) has been a boon to the world of scientific publication. New research papers can be disseminated more quickly and for less cost than ever before, resulting in a tremendous increase in the quantity and diversity of easily available research publications. However, this has exacerbated the problems of information overload for researchers attempting to keep abreast of new relevant research, especially in rapidly advancing fields.

Scientific literature accessible through the Web can be treated as a massive, noisy, disorganized database from which researchers would like to extract knowledge about important new developments, and be able to track new research trends. However, unlike many large “single source” databases (e.g. a corporate customer database), the research publications on the Web come from a large number of sources, each of which may have its own organization. Also, the diversity of research topics on the Web means most of the records in such a data set are irrelevant. Furthermore, the database is constantly growing and changing in both composition and organization. This lack of regular organization, high degree of inclusion of unimportant records, and highly dynamic nature make this a particularly difficult domain for knowledge discovery.

In order to automatically discover useful knowledge from such a database, it may be important to include a system of information filtering (IF). IF is the process of extracting *only relevant* records as they appear in a stream of new incoming records. (See [1] for an introduction to IF). Thus, the problem of finding important new research requires filtering to extract only publications that may be relevant or interesting to the user, as well as extracting specific publications, concepts, and trends that may indicate important new research developments. To this end, we have developed the *CiteSeer* digital library system [2].

CiteSeer is a generator of custom digital libraries that performs several information filtering and knowledge discovery functions in order to keep users up to date with the latest relevant research. CiteSeer's knowledge discovery process is comprised of three major components: (i) Database creation and feature extraction, (ii) personalized filtering of new publications, and (iii) personalized adaptation and discovery of interesting research and trends. These parts are interdependent in the sense that the information filtering affects what is discovered, and good discoveries (as judged by the user) are used to tune the information filtering process.

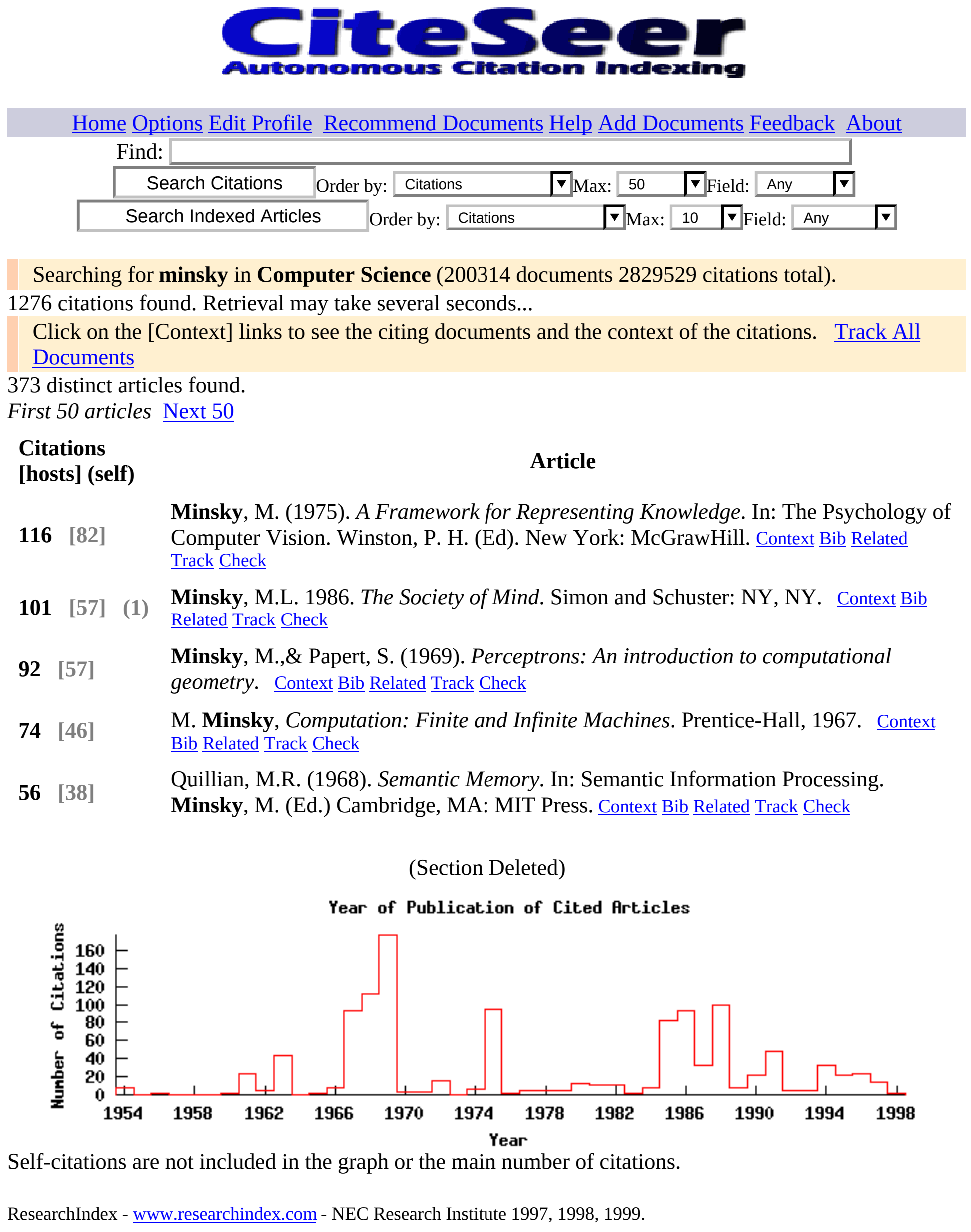
2 Database Creation and Feature Extraction

The body of scientific literature on the Web is spread among many Web sites, is usually in an unsearchable form (e.g. Postscript or PDF), and is organized differently at each Web site. CiteSeer creates a database by downloading publications from the Web in a general area of research, such as neural networks, or computer vision. This first stage is static and performed by heuristics. Once downloaded, CiteSeer extracts the raw text and parses it to extract various fields common to most research papers, such as title, abstract, word frequencies, and list of citations. These features are indexed and placed in a local database.

Rather than utilizing simple template matching, CiteSeer uses sophisticated heuristics to parse a wide variety of paper formats. For example, the title of a paper can be identified almost always by finding the largest font on the first page. Also, citations to the same paper may be in different formats, depending on the citing paper, so clustering techniques are used to identify these as being the same citation with high reliability [3]. Since both a paper and citations to that paper may be in the database, matching of title and use of other heuristics can be used to automatically tie a paper to citations of that paper. This allows us to then build the full graph of citing and cited papers.

At this stage, CiteSeer provides a variety of static searching and browsing capabilities to greatly reduce the amount of effort required to perform a literature survey. Beyond traditional keyword search on the paper text and citations, CiteSeer provides the facilities to browse forward and backward through citation links, allowing both citing and cited papers of an interesting work to be found. It extracts and summarizes citation contexts to make quick appraisal of papers easier, and gives citation statistics including the number of citations for each cited paper and identification of self-citations.

The details of these capabilities in CiteSeer may be found in [4]. It has a Web browser based interface from which users may perform searches on the downloaded documents. For example, consider the search for citations of the author "Minsky" as shown in Figure 1. This query was



Home Options Edit Profile Recommend Documents Help Add Documents Feedback About

Find:

Search Citations Order by: Citations Max: 50 Field: Any

Search Indexed Articles 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

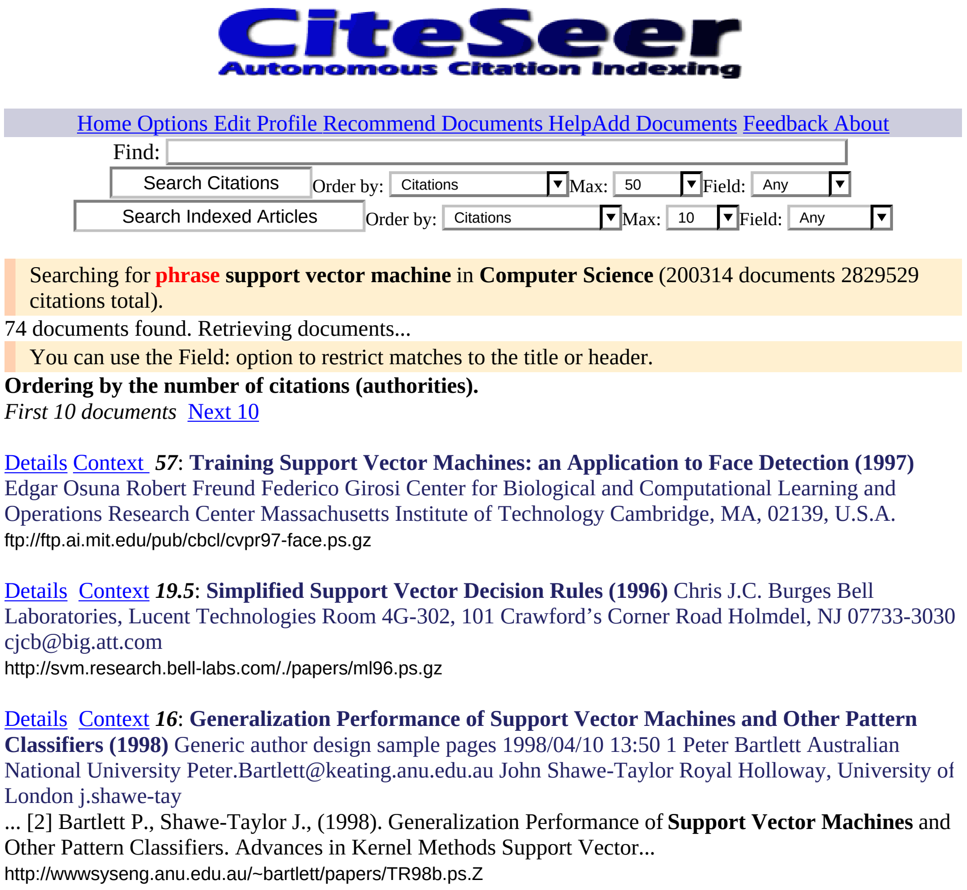
(Section Deleted)

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: The results of a CiteSeer query for citations of “Minsky”.

performed on a small database of computer science papers (about 200,000 documents having 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 a user's article query of "support vector machine" with a list of papers ranked by the number of times they are cited in the database, as shown in Figure 2. If the user is especially 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. The first part of these details are shown in Figure 3.


CiteSeer
Autonomous Citation Indexing
Home Options Edit Profile Recommend Documents HelpAdd Documents Feedback About
Find:
Search Citations
Search Indexed Articles

Order by: Citations Max: 50 Field: Any

Order by: Citations Max: 10 Field: Any

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

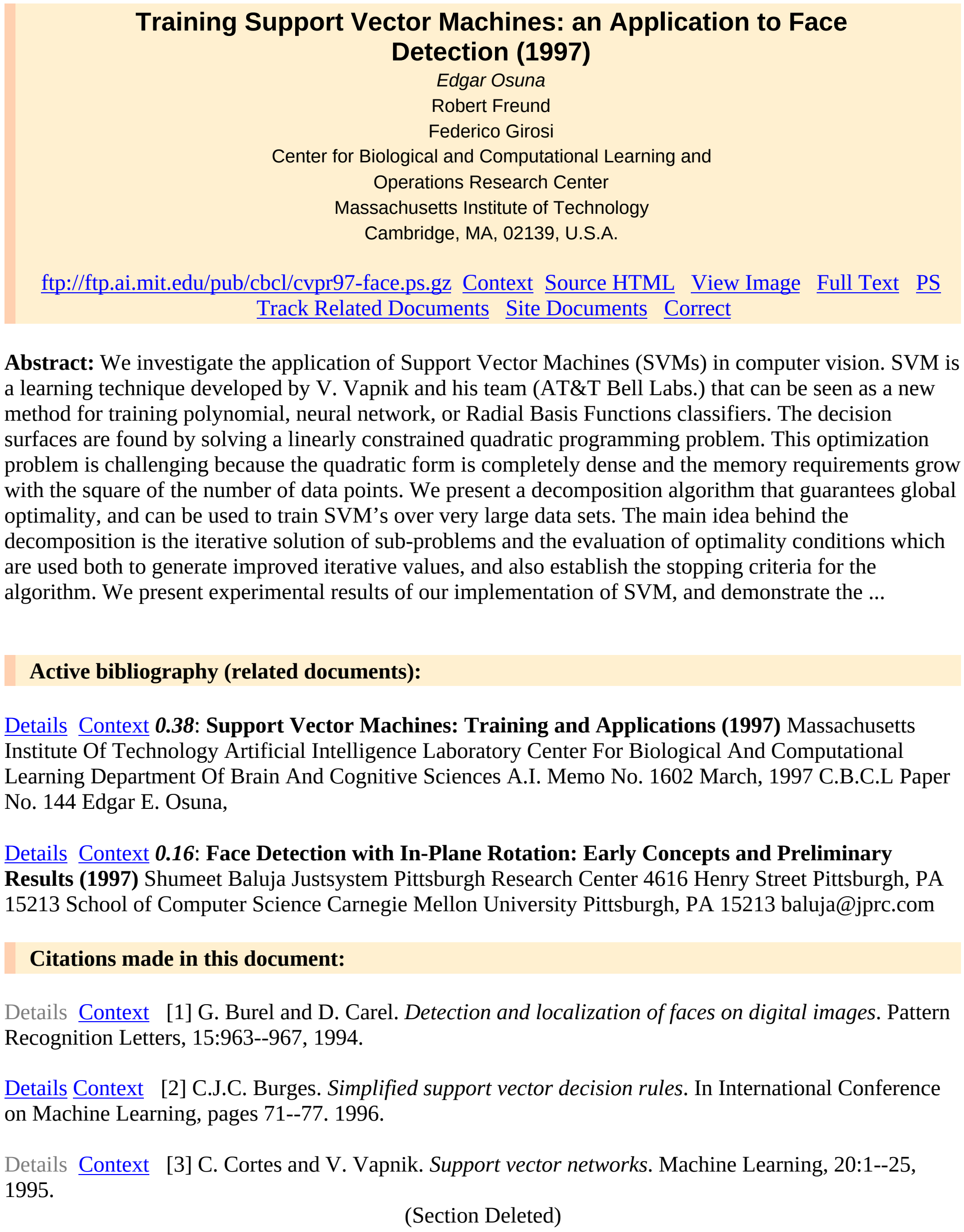
Details Context ***19.5***: **Simplified Support Vector Decision Rules (1996)** Chris J.C. Burges Bell Laboratories, Lucent Technologies Room 4G-302, 101 Crawford's Corner Road Holmdel, NJ 07733-3030 cjcb@big.att.com
http://svm.research.bell-labs.com/./papers/ml96.ps.gz

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)

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

This first part of the CiteSeer system extracts features from a disorganized and essentially unsearchable source (Postscript or PDF documents on the Web) to build a digital library and provides useful tools for finding relevant scientific literature in this library. This feature uses several heuristics so as to be well tuned to the structured internal organization of scientific literature, and sets the stage for more sophisticated adaptive filtering and discovery.



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 3: The top part of document details of the paper "Training Support Vector Machines: an Application to Face Detection".

3 Personalized Filtering

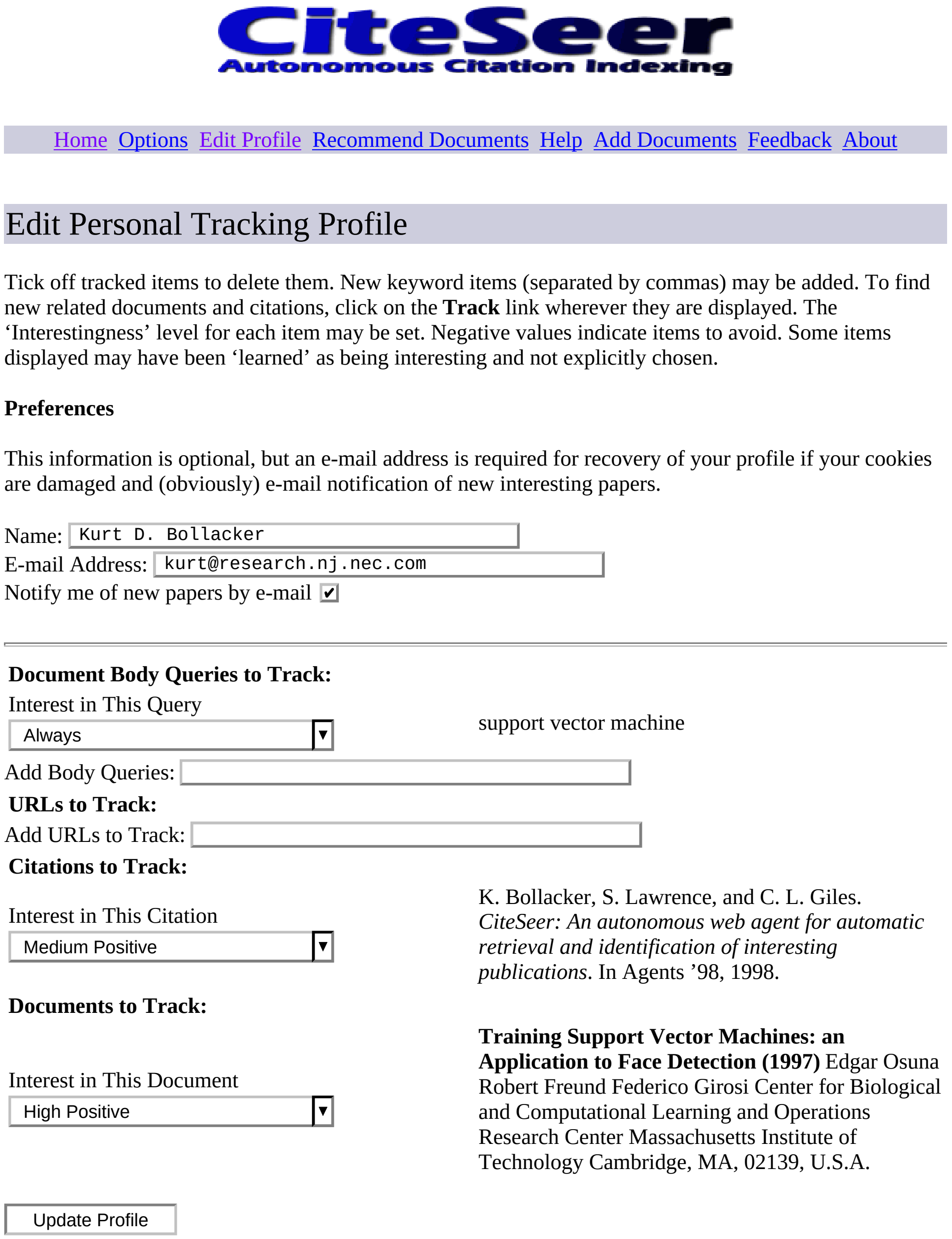
CiteSeer uses personal profiles representing a user's research interests to track and recommend new relevant research. CiteSeer can examine the local database of publications to determine whether any new papers may be considered interesting by the user and alert the user by e-mail or through a Web based interface. The profile is adaptive to the user's research interests by a system of feedback using manual profile adjustment and machine learning. CiteSeer watches the user's browsing behavior and response to recommendations made by CiteSeer to modify the user's profile. These modifications may result in new recommendations, to which the user again responds. Over time, this cycle of learning may allow CiteSeer to find relevant papers with better accuracy and reliability.

The information provided by a user in this learning regime consists of both explicit modifications to the profile and implicit opinions gathered from user actions during browsing or responding to recommendations. When the first bit of information appears, a new profile is created and CiteSeer begins to learn.

3.1 Profile Creation

In the process of using CiteSeer's Web interface, a user contributes to his or her profile explicitly by manual editing of the profile or implicitly by browsing a database. Either of these actions create or modify profile components we call *pseudo-documents* which are used as the representation of the user's research interests. Pseudo-documents are place holders for a set of values of (often only single or few) features extracted from publications. The question of which features to extract to form a pseudo-document is an area of active research (e.g. [5, 6]). CiteSeer uses a heterogeneous set of pseudo-documents including features such as keywords, URLs, citations, word vectors and citation vectors. There is evidence that this may be more powerful than any single representation [7, 8]. For example, [9] shows that retrieval of papers based on citation features has little overlap with retrieval based on keywords. Thus, a user's profile consists of a set $\mathcal{D}$ of different types of pseudo-documents. In addition to a feature value, each pseudo-document d has a weight w_d which corresponds to its influence. For example, high positive w_d values mean the pseudo-document is a very good example of what the user is interested in, and a negative value indicates an item the user would like to avoid.

CiteSeer's facility for users to explicitly create a profile from a Web interface is shown in Figure 4. From this Web page, users may add or modify the influence of keyword or URL feature values for constraint matching, or may modify the influence of citations or relevant papers previously specified in the process of browsing CiteSeer. The influence of each item may be manually adjusted. For the example profile shown here, the user has selected the **Track Related Documents** link from Figure 3.



CiteSeer
Autonomous Citation Indexing

Home Options Edit Profile Recommend Documents Help Add Documents Feedback About

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 4: A sample CiteSeer user profile. The user can create and modify the influence of components in order that the profile better reflects the user’s interests.

3.2 Interestingness of New Papers

When a new paper d^* becomes available on the Web and is added to the database, it is treated as a pseudo-document having features corresponding to the union of the feature types in the user's profile $\mathcal{D}$. This pseudo-document is compared to those in the profile, to give a level of similarity $I_{\mathcal{D}}(d^*)$ interpreted as the "interestingness" of the new paper. More specifically, this is calculated as the weighted sum:

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

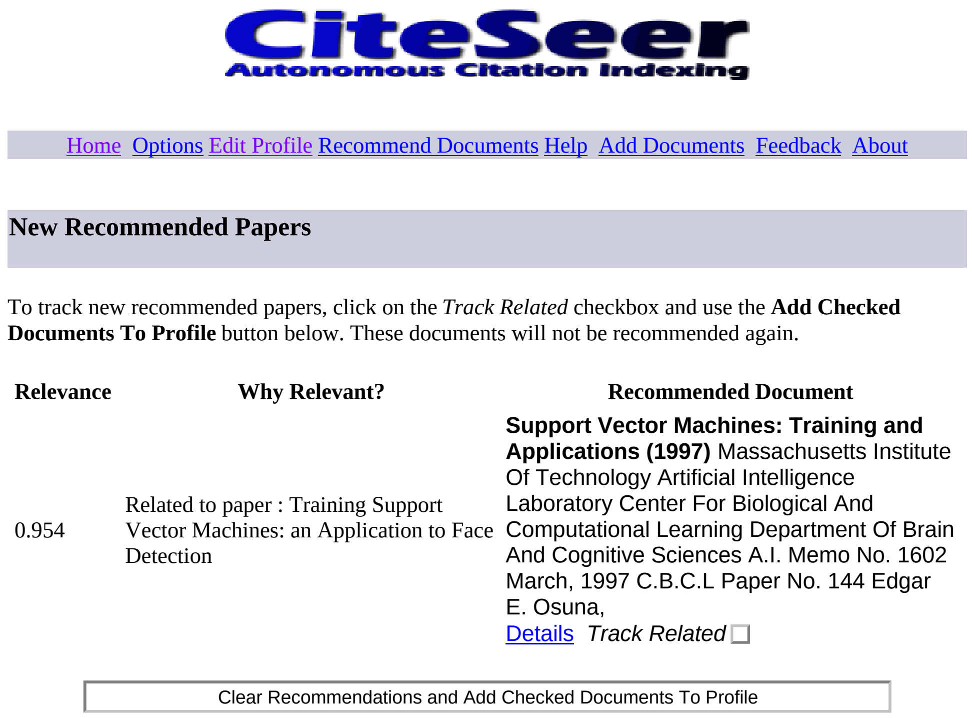
where $R_d(d, d^*)$ is the similarity (or relatedness) between a pseudo-document d in the user's profile and the new paper pseudo-document. Each relatedness measure is weighted by the profile pseudo-document's influence. Those new papers d^* which have a $I_{\mathcal{D}}(d^*)$ greater than a threshold are recommended to the user. Currently this is set to be a small positive number, but is effectively adjusted on a per user basis as described later.

The measure $R_d(\cdot, \cdot)$ used to determine relatedness is dependent on the type of pseudo-documents. For example, the user can create pseudo-documents as explicitly specified keywords, citations, and other constraint values. In the case of a constraint, the appropriate relatedness measure is currently a simple zero or one depending on whether the new paper matches that constraint. Although constraint based similarity is useful, often a user would like to find papers which are related even if they do not match any given constraint. The user would like to simply say, "Tell me about new papers that are related to these existing papers."

One such measure that captures this idea of relatedness between two papers we call Common Citation × Inverse Document Frequency (CCIDF) [4], and is defined as the sum of the inverse frequencies of the common citations between the two papers. If two scientific papers cite many of the same previous publications, then these two papers are likely to be related. Furthermore, if a cited work is very obscure, then this is a more powerful indicator of relatedness than one well known and often cited. CCIDF is similar to the notion of bibliographic coupling, and is partially analogous to the word vector based TFIDF [10]. *(See the sidebar for a discussion of various relatedness measures.)*

3.3 Paper Recommendation

Once a profile has been created, periodically (or on demand) CiteSeer will check to see if any new papers that should be recommended to the user have been added to its database. These recommendations are sent via e-mail if desired, and are also presented when the **Recommend Documents** link is chosen as shown in Figure 5. Papers are ranked by their $I_{\mathcal{D}}(\cdot)$ value and an explanation for recommendation is given. This explanation is simply a listing of type and value of the pseudo-document in the profile contributing the most to the recommended paper's interestingness. Any of the recommended papers can be viewed, downloaded, ignored, or explicitly added to the user's profile.


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

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 5: A new paper found by CiteSeer and recommended to the user as potentially interesting. Recommendations include the $I_{\mathcal{D}}(\cdot)$ value and an explanation of why the papers were recommended.

3.4 Profile Adaptivity

CiteSeer adapts a user's profile to better represent his or her interests by the modification of the pseudo-document weights w_d. CiteSeer does this in three major ways: (i) observing user behavior during database browsing, (ii) allowing manual adjustment, and (iii) learning from user responses to recommendations. CiteSeer observes the use of its Web interface to modify the profile of new papers. There are several types of user actions that can be observed and used as implicit indications of interest in the (pseudo-document) object of that action[11]. We have chosen several of these: viewing details, downloading a paper, explicitly adding/removing a paper to/from the profile. For example, if the user chooses to view the details of a paper (as in Figure 3) a CCIDF pseudo-document for this paper is added to (or modified in) the user's profile. Each type of user action b_d on pseudo-document d initializes or adds to the influence w_d of d an amount corresponding to the interestingness $a(b_d)$ indicated by that action as given in Table 1. These values are currently set in an ad hoc manner and are fixed in the current CiteSeer implementation. The special case of explicitly adding to or modifying pseudo-documents an be considered to be manual adjustment of the profile. However, manual adjustments to the w_d values are also allowed.

When a document d^* is recommended, CiteSeer observes the user's response to the recommendation and updates the weight for each pseudo-document d in the profile $\mathcal{D}$ accordingly. The update rule is

$$w_d \Leftarrow w_d + \eta a(b_{d^*}) R_d(d^*, d) \tag{1}$$

User Action b	Document Interestingness $a(\cdot)$
Paper Explicitly Added To Profile	Very High Positive
Paper Downloaded	High Positive
Paper Details Viewed	Moderate Positive
Paper Ignored	Low Negative
Paper Removed From Profile	Set To Zero

Table 1: The "interestingness" of a paper as determined by user actions on that paper.

where η is a learning rate, and $R_d(d^*, d)$ is the relatedness measure for the specific type of pseudo-document d. Although simple, the update rule of Equation 1 has several useful properties:

- Weights on pseudo-documents that contribute to good recommendations are increased while weights on pseudo-documents that contribute to bad recommendations are decreased.
- The overall precision/recall threshold of the system is implicitly and automatically adapted. If the threshold is too low, then too many documents of poor relevance will be recommended and ignored, thus lowering the w_d values, which effectively raises the threshold. If the threshold is too high, then too few documents will be recommended, and the user is thus encouraged to add more pseudo-documents.
- The influence of different relatedness measures is adapted separately. This allows real documents in the profile that are interesting in only some ways to be used to find good candidate documents using only those ways. Relatedness measures that are poorly correlated with $a(\cdot)$ will (hopefully) tend to have little influence.
- Both explicit and implicit feedback from the user are utilized. This is the best of both worlds, because although explicit feedback is much easier to use and tends to be more accurate, it is much harder to acquire.
- The model is computationally scalable. The cost of interestingness calculations and profile updates are linear with the size of the profile, and do not increase with the size of the database.
- New relatedness measures and corresponding pseudo-document types can be easily added.

4 Personal Knowledge Discovery

CiteSeer's system of profile adaptivity through manual adjustment and machine learning can provide more than simply a way to find better papers to recommend. Once a profile has been well tuned to a user's interests, it is possible to apply discovery techniques to find new research concepts and trends that may be of interest to the user. Also, some of this discovery may happen simply as a result of CiteSeer's functionality.

4.1 New Concepts

CiteSeer tends to increase the weights of pseudo-documents that contribute greatly and/or often to good recommendations. Correlations between these highly weighted pseudo-document values and other feature values extracted from the same papers can reveal interesting new concepts. For example, author names which correlate highly with citations made by papers in the user's profile, but are not already part of a constraint based pseudo-document, may be suggested.

4.2 Changes With Time

Over time, a user's interests may change and grow, requiring more rapid and substantial updates to the user's profile than initial tuning to a specific interest. If a user adds new papers to his or her profile from a new research area, these papers may not be substantially related to existing papers in the profile. In this case, CiteSeer's design automatically maintains multiple "interest clusters" without any explicit consideration of such. Discovery of such clusters should be possible using traditional clustering techniques.

4.3 New Areas of Research

If papers appear from a new research area that is potentially interesting to a user, the user may not have authors or keywords in his or her profile to discover these new papers. However, these papers must cite previously published research. If some of these citations are to papers that are part of the user's profile or result in high enough relatedness to papers in the profile, CiteSeer may recommend these new papers. This demonstrates how citation based features can be instrumental in discovering new, interesting research trends.

5 Conclusions and Future Work

CiteSeer currently is a system of many information extraction, adaptation, and knowledge discovery tools that allow users to keep up to date with new published research on the Web that is related to their interests. Although the CiteSeer system informally seems to be very useful, there is much work to be done in the future. Evaluation of how well the CiteSeer profiles represent and learn changes in user interests needs to be performed. This may include techniques such as cross validation using random partitioning of the profile into training and test sets of pseudo-documents. Also, we intend to explore more sophisticated analysis and knowledge discovery techniques to allow better identification of personally important research trends. For example, a CiteSeer database may be treated as a directed graph where citations are edges and papers are nodes. Citation graph analysis may result in better relatedness measures or discovery of structural features such as citation cliques by mapping to an author citation graph. Also, other technologies such as collaborative filtering may increase CiteSeer's power to find new, interesting papers that would otherwise be missed.

A demonstration CiteSeer database of over 200,000 computer science research papers having about 2.8 million citations is currently available at **http://csindex.com**. Readers are encouraged to use this free, publicly available service and provide feedback.

Acknowledgements: The authors would like to thank Eric Glover, Gary Flake, and Nelson Amaral for their helpful suggestions and comments.

6 SIDEBAR – Relatedness Measures Between Research Papers

6.1 Document Relatedness Measures

When a new candidate paper appears, CiteSeer must decide whether to recommend the paper to the user. If a paper has been determined to be sufficiently similar to the collection of pseudo-documents making up a user's profile, then it is considered to be related, and is recommended. Generally, let $R_d(d, d^*)$ be a relatedness measure between a pseudo-document d that is appropriate for d's type of pseudo-document and the new candidate document d^*. Each type of relatedness measure is specific to the type of pseudo-document for which it is used.

6.2 Constraint Based Relatedness

Constraint based relatedness is generally used with pseudo-documents in the profile that are not part of a real paper. For example, a user may specify the term "support vector machine" to be a desirable keyword. The pseudo-document d that represents this is an artificial document that has as its only feature this keyword. If a candidate document d^* contains this keyword, then the relatedness $R_d(d, d^*)$ is unity and zero otherwise.

6.3 TFIDF Word Vector Relatedness

It is common to consider a document to be a collection of words upon which statistics can be gathered. The frequency of each unique word stem can be measured. A feature vector $\vec{W_D}$ is extracted and used as a pseudo-document d where each component is the frequency of the word stem in the document. One often used form of this measure is known as *term frequency* × *inverse document frequency* (TFIDF) [10]. In this scheme, the feature set $\vec{W_D}$ is a vector of word frequencies weighted by their rarity over a collection of documents. Let $\mathcal{W}$ be the set of all unique words in the CiteSeer database. In a pseudo-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 such TFIDF scheme [12] a word weight vector element w_{ds} is calculated as:

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

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

$$R_d(d, d^*) = \vec{W_D} \cdot \vec{W_{D^*}}. \tag{3}$$

6.4 Citation Based Relatedness

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, this is a more powerful indicator than if a citation is to an extremely well known and often cited publication. A measure that captures this idea of relatedness we call "Common Citation × Inverse Document Frequency" (CCIDF) [4] and is partially analogous to the word vector based TFIDF. Let f_i be the frequency of a citation i in the CiteSeer database, let $C_i = 1/f_i$ be the inverse frequency, and let $\vec{\mathcal{C}}$ be the vector of these inverse frequencies. Let c_{di} be a Boolean indicator of whether pseudo-document d contains citation i and let $\vec{X_d}$ be the resulting Boolean vector. The CCIDF relatedness between a candidate pseudo-document d^* and a pseudo-document d in the profile is defined as:

$$R_d(d, d^*) = tr(\vec{X_d} \times \vec{X_{d^*}}) \cdot \vec{\mathcal{C}}$$

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

References

[1] C. Faloutsos and D. Oard, "A survey of information retrieval and filtering methods." University of Maryland Technical Report CS-TR-3514.

[2] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," *IEEE Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[3] S. Lawrence, K. Bollacker, and C. L. Giles, "Autonomous citation matching," in *Proceedings of the Third International Conference on Autonomous Agents* (O. Etzioni, ed.), (New York), ACM Press, 1999.

[4] K. Bollacker, S. Lawrence, and C. L. Giles, "CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications," in *Proceedings of the Second International Conference on Autonomous Agents* (K. P. Sycara and M. Wooldridge, eds.), (New York), pp. 116–123, ACM Press, 1998.

[5] E. Bloedorn, I. Mani, and T. R. MacMillan, "Representational issues in machine learning of user profiles," in *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, March 1996.

[6] B. Krulwich and C. Burkey, "Learning user information interests through extraction of semantically significant phrases," in *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, March 1996.

[7] M. Balabanovic, “An adaptive Web page recommendation service,” in *Proceedings of the First International Conference on Autonomous Agents*, February 1997.

[8] B. T. Bartell, G. W. Cottrell, and R. K. Belew, “Automatic combination of multiple ranked retrieval systems,” in *Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*, pp. 173–181, 1994.

[9] K. McCain, “Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution,” *J. Amer. Soc. Inform. Sci.*, vol. 40, no. 2, pp. 110–114, 1989.

[10] G. Salton and C. Yang, “On the specification of term values in automatic indexing,” *Journal of Documentation*, vol. 29, pp. 351–372, April 1973.

[11] D. M. Nichols, “Implicit rating and filtering,” in *Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pp. 31–36, November 1997.

[12] G. Salton and C. Buckley, “Term weighting approaches in automatic text retrieval.” Tech Report 87-881 Dept. of Computer Science, Cornell University, 1997.