

Customization Using Support Vector Machines for Information Retrieval

Kenro AIHARA and Atsuhiko TAKASU

National Institute of Informatics

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

{kenro.aihara,takasu}@nii.ac.jp

ABSTRACT

This paper proposes an HCI approach to supporting interactive document retrieval in unorganized open information space.

On the assumption that taxonomical thought are one of the most important and skilled operations for us when we organize or store information, we may discuss that clustering or classification techniques are applied to interactive interface. The proposed system, therefore, provides two-dimensional space to visualize categories of document icons for interactive interface. The features of the proposed approach are (1) visualization of document categories for interaction, (2) initialization of categories by hierarchical clustering method, (3) customization of categories by support vector machine techniques, (4) user-depend additional attributes for individual implicit, and (5) cognitive aspects integration of these techniques for IR interface.

We made a preliminary experiment on text categorization for customization, using two test collections to estimate practicality.

Keywords: User Feedback, Customization, Visualization, Human-Computer Interaction, Text Categorization, Support Vector Machine, Information Retrieval

1. INTRODUCTION

Human-computer interaction (HCI) has become one of the key issues for information retrieval (IR) systems. Many of existing IR systems require its user's declaration of keywords related to his/her information needs or the range of attribute values to support the user to access the required information. However, it is difficult to externalize users' requirement with necessary and sufficient words. In addition, it is not easy to recognize or evaluate the answered results. Usually there is a gap of recognition between information providers, such authors or database designers, and the users.

In recent years, on the other hand, information space which we can access via Internet has been getting bigger and bigger. The augmentation has brought to us a new problem, information flood; how to know where what I want exists. Recently numerous studies have been made on mediating between information sources.

This paper proposes an HCI methodology to supporting interactive document retrieval in unorganized open information space. The features of the proposed methodology are as follows:

1. visualization of document categories for interaction
2. initialization of categories by a hierarchical clustering method

3. customization of categories by the support vector machine technique

4. user-depend additional attributes for individual implicit cognitive aspects

5. integration of these techniques for IR interface.

Section 2 describes background of this paper. Section 3 illustrates our basic idea and scenario of our system usage. Section 4 explains our user feedback approach. Section 5 shows a preliminary experiment. At last, Section 6 summarizes this paper.

2. BACKGROUND

Studies on data mining and knowledge discovery from databases approach information flood problem from information sources' side [12, 9]. In these studies, implemented systems try to extract data ordinalities or relativities. Recently numerous studies have been made on mediating between information sources. Agent-based approach [4, 22], integration of schemes of databases [14] are currently hot issues.

Studies on HCI for IR approach from the users' side. Visualization of information space helps users to recognize features of information or relations among objects. Many studies focus on visualization from geometrical aspects [6], for example, visualization of hyperlink structure or history of the users' references of World Wide Web (WWW) as a graph [15]. However, it should be noticed that we have to consider what and how to visualize so that IR system become effective for the user. Some studies deal with content-depend visualization [16, 11, 7, 10, 24]. In these visualization approaches, a target of document set gets restricted at first. Then a system tries to let its users notice effective terms through a cycle of visualizing icons of the documents, definition of region of the target documents, or selection of terms. We, however, cannot believe that a directly application of this approach to large-scale document sets would be effective.

Information filtering or social filtering studies are application techniques to select preferable information from information flow [13, 17, 20]. These studies attempt to apply user preferences or community factor to IR techniques.

Text categorization studies are getting more important recently because WWW has let us notice the necessity and interest of global analysis. Many categorization methods have been proposed: Rocchio, naive Bayes, latent semantic indexing, decision theory, nearest neighbor, neural net, or SVM [27, 19]. These studies are facilitated by using recent rich computational resources and test collections provided for research purpose.

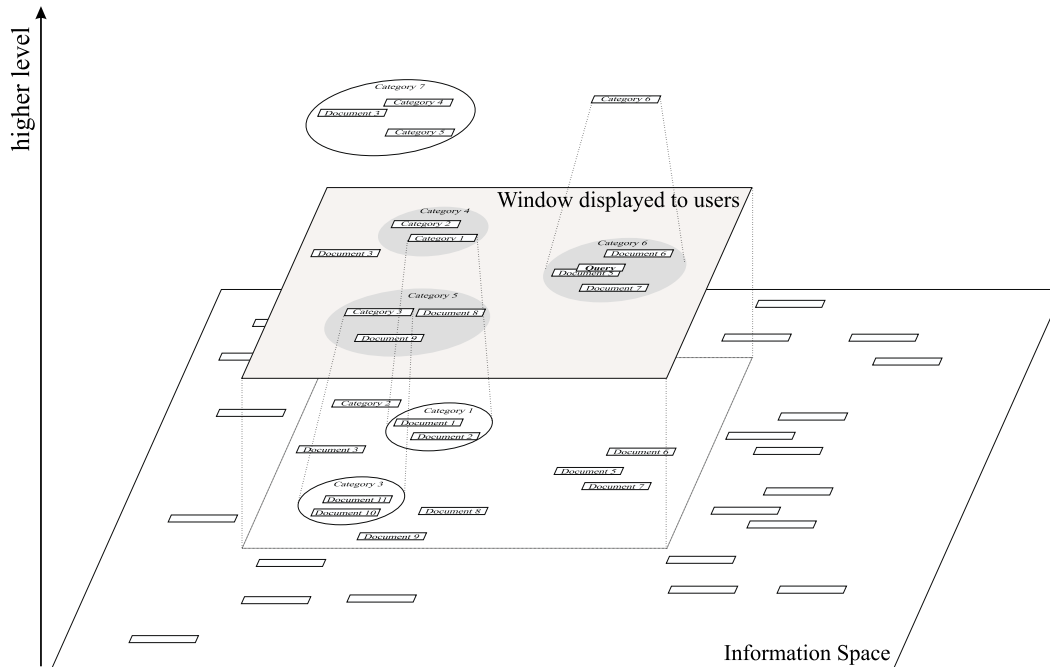


Figure 1: Arrangement of Documents in 2-dimensional Space and Hierarchical Categories

3. METHODOLOGY

Basic Idea

When information systems attempt to support its users to get information about their needs, there can be the following two aspects:

Unification of Information Space construction of global scheme and common terminology, or ontology, and instruction of them to all users

User-centered Information Access extracting users' information needs through interaction with information space and providing information according to users' implicit dimensions

This paper focuses mainly on the latter one.

We suppose that it is effective that the system let users recognize the information space inclusively and an effective interaction system should have the following features:

- visualization for users' easy and intuitive recognition of information
- customizability through natural interaction (without many examinations of initial retrieval)

These features must be needed for usability related to users' load and effectivity for their primary retrieval purpose.

We have studied human-computer interaction to access information sources. Our research has been focused on categorizing documents for visualization, and it is related to the former feature.

User relevance feedback, on the other hand, is the most popular approach to help users to formulate queries to retrieve expected information for each user through interaction between him/her and the system. Although this approach deals with the latter

feature, a great deal of proposed methodologies make its user examine ranked documents in every retrieval task. We should discuss another approach which needs less examination load for users.

We guess that taxonomical thought are one of the most important and skilled operations for us when we organize or store information. The way of classification may change dynamically according to one's mental and physical situation. On that assumption, we may discuss that clustering or classification techniques are applied to interactive interface.

The proposed system, therefore, provides two-dimensional space to visualize information space. Information objects, such as documents, are placed in the space as icon.

In addition, the system also manages the hierarchy of categories. Hierarchical categories can keep the number of items or concepts which users must handle at once small.

Figure 1 illustrates categorization of documents and an hierarchical categories of documents of our approach. The system shows a rectangle area (represented by "Window displayed to users" in the figure) of the information space including all documents at a height level specified by the user. The arrangement of icons are computed based on a kind of multi-dimensional scaling method as "similar" documents should be placed close in the space.

Interactive Scenario

In this section, we show an interactive scenario of the proposed system. Figure 2 illustrates the system architecture.

At first, the system stores "the document-term matrix" which has entries a weight $f(i, j)$ of term j for document i from a collection.

Then the system produces initial categories of documents, "Category Information", by a clustering method.

Next, the system shows document icons in two-dimensional

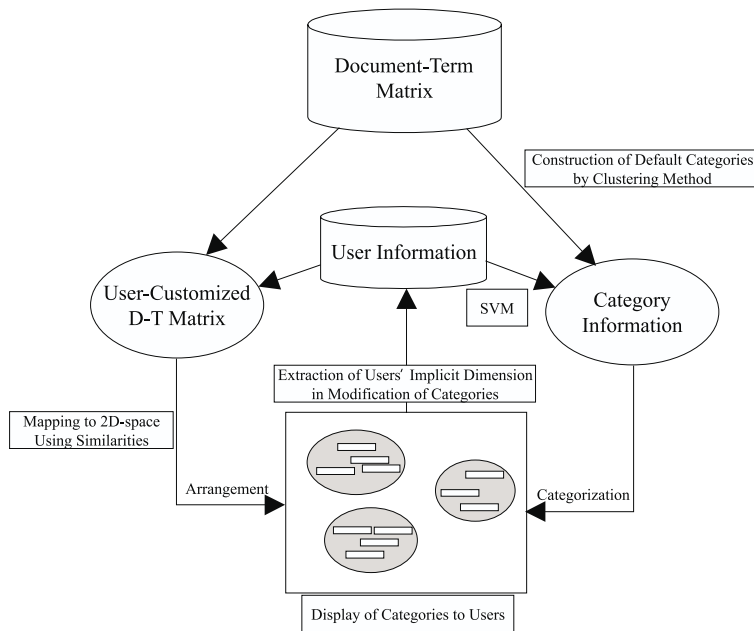


Figure 2: System Architecture

space as shown in Figure 1. Arrangement of icons are computed using similarities of documents.

When a user views this arrangement of documents, he/she may recognize a structure of the information space, relation among documents, or any features of documents. He/she may find out his/her well-known documents. The user may also predict the content of a document which placed proximal zone of such well-known documents.

From previous researches, we know that the user cannot agree the shown categories. When the user has an unagreeableness in his/her mind, he/she often tries to modify such visualized information as he/she likes. We suppose that that action must include the user's implicit cognition, such an intent which cannot be verbalized by him/herself. The system, therefore, can exploit such actions as the feedback information which is stored in "User Information" of Figure 2.

For customization, user feedback is applied to categorization of documents and their arrangement. Thus user feedback information is stored for each user and the system gets customized.

When its user wants to retrieve information or browse the information space, he/she can use his/her this own customized IR interface. If a user inputs a query by keywords or an example document, its query vector or the example document would be mapped into an appropriate category. Of course the user can also browse the documents without query.

Besides, newcoming documents will be classified into an appropriate category. This approach, therefore, can be used for information filtering system.

4. User Feedback

The system can be customized in user feedback process. The feedback will be given by users' modification of categories which shown in the system (Figure 3).

If document d , such "Document 4" in Figure 3, (or cluster of

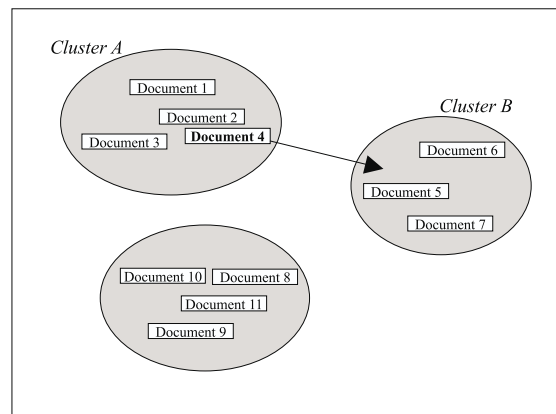


Figure 3: User Feedback in Modification of Categories

documents) in cluster A is moved into cluster B by a user, the system appends a new attribute for the user. In our research[2] or many relevance feedback studies, systems usually process feedback in given dimensions. We, however, suppose that such original dimensions couldn't represent the user's feedback well enough. For extrapolation of insufficient dimension related to users' cognitive factor, the system appends user attributes.

In fact, categories of documents and arrangement of icons in the 2D space are customized in feedback process. Customization processes are presented in the sections below.

Support Vector Machine

This section illustrates support vector machine (SVM) [5] which is the key technology of customization of categories.

SVM is a classifier for multi-dimensional data to determine a boundary curve between two classes. The boundary can be determined only with vectors in boundary region (so-called "mar-

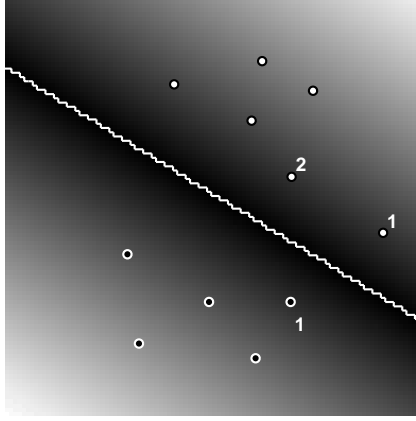


Figure 4: Support Vector Machine (from [5])

gin”) of two classes (so-called “support vectors”) in a training examples. SVM, therefore, need to be re-learned only when vectors in boundary change.

From the training examples SVM finds the parameters of the decision function $D(\vec{x})$ which can classify two classes A and B and maximize margin during a learning phase. After learning, the classification of unknown patterns is predicted according to the following rule:

$$\begin{aligned} \vec{x} &\in A && \text{if } D(\vec{x}) > 0 \\ \vec{x} &\in B && \text{otherwise} \end{aligned}$$

Figure 4 illustrates SVM. The gray levels encode the absolute value of the decision function (solid black corresponds to $D(\vec{x}) = 0$). The numbers indicate the supporting vectors.

SVM has the following advantages:

- the solution is unique
- the boundary can be determined only by its support vectors, namely SVM is robust against changes of all vectors but its support vectors
- SVM is insensitive to small changes of the parameters
- different SV classifiers constructed by using different kernels (polynomial, RBF, neural net) extract the same support vectors[23]

The second advantage is very important for a large-scale and open document collection which change constantly.

Strategies based on reference vector which is produced from all vectors categorized to a class, such an average vector, need to re-compute the reference vector when a vector which belongs to the class changes. In general, reference vector strategy should be sensitive to changes of vectors. Figure 5 illustrates boundaries of both strategies. The left figure shows an example of reference vector strategies, and the other indicates one of support vector strategy. If a vector changes, reference vectors (designated by crosses in the left figure) which contain the changed vector would change and that cause the change of boundaries. On the other hand, SVM need to re-learn only when support vectors (designated by crosses in the right figure) changes.

For multi-class classification, we use as many SVMs as the number of classes because SVM classifies vectors into two classes.

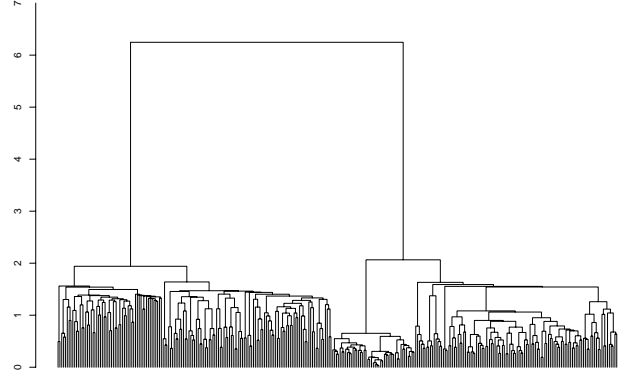


Figure 6: An Example of an Hierarchical Clustering of Documents (technical papers and articles of Journal of Japanese Society for Artificial Intelligence)

Initial Categories by Clustering

The proposed IR interface shows document categories to users. We expect that visualized document categories can help users to recognize the shown information space regardless of its user’s familiarity of the domain of the information.

Initial categories can be obtained by clustering.

Many clustering methods, such hierarchical clustering, graph based clustering, or neural network based clustering, have been proposed. Willet reviewed clustering methods in IR studies [26].

We use a hierarchical clustering and the complete link algorithm. A hierarchical clustering can divide all documents into some sub-clusters according to its dendrogram and sub-cluster can be represented as reference documents. The complete link algorithm computes an inter-cluster similarity as the minimum of the similarities between all pairs of inter-cluster documents. The algorithm merges the pair of clusters with the highest inter-cluster similarity. It, therefore, can produce small and tight clusters.

Figure 6 shows a dendrogram of hierarchical clustering. This hierarchy consists of text corpus of technical papers and articles of Journal of Japanese Society for Artificial Intelligence. This corpus consists of 674 documents published from 1986 to 1995.

We have to discuss further on which clustering method is the most effective or useful.

Customization of Categories by SVM

After a user modifies categories, the system re-learns the new boundary of cluster B, B-against-the-rest boundary, by SVM with the incoming documents and support vectors.

As mentioned above, the system appends user attribute after modification. For learning of SVM, the system sets 1 to the value of the attribute of moved document d (or documents which belong to the moved category). This attribute can be used for distinguishing between the moved documents and ones in the previous category. It might be difficult to classify them without this appended attribute. And SVM learns the ordinality from the original attributes within the new category.

The user-depend weight for categorization $f_{cat}(u, i, j)$ of attribute j of document i for the user u is given by Eq. (1) and

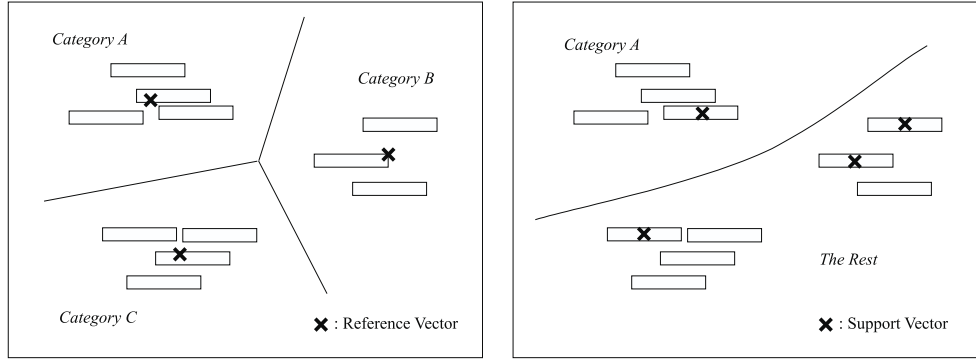


Figure 5: Reference Vector vs. Support Vector

Eq. (2).

$$f_{cat}(u, i, j) = \begin{cases} f(i, j) & (j = 1, 2, \dots, m) \\ U_{cat}(u, i, k) & (j = m + k, \\ & k = 1, 2, \dots, N_a(u)) \end{cases} \quad (1)$$

$$U_{cat}(u, i, k) = \begin{cases} 1 & (\text{if } i \text{ was modified} \\ & \text{in } k\text{-th action}) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

where m stands for the number of terms and $N_a(u)$ stands for the number of feedback actions of the user u . Thus categories can be customized with user attributes and SVM.

The system needs to re-learn not all SVMs of categories but only one of the target category when a new attribute is appended. That is very important advantage of SVM. The advantage enables us to append attributes in actuality.

Customization of Arrangement

In contrast with learning of SVM, the user-depend weight for arrangement $f_{arr}(u, i, j)$ is given by Eq. (3) and Eq. (4).

$$f_{arr}(u, i, j) = \begin{cases} f(i, j) & (j = 1, 2, \dots, m) \\ U_{arr}(u, i, k) & (j = m + k, \\ & k = 1, 2, \dots, N_a(u)) \end{cases} \quad (3)$$

$$U_{arr}(u, i, k) = \begin{cases} 1 & (\text{if the category of } i \text{ was} \\ & \text{modified in } k\text{-th action}) \\ 0 & (\text{otherwise}) \end{cases} \quad (4)$$

The system sets 1 to the value of the appended attribute of all documents which belong to the target category of the moved document for arrangement of documents. The reason is that documents which the user regards as similar should be placed close and the similarity between them must be higher.

5. EXPERIMENTAL RESULTS

We made a preliminary experiment on text categorization for customization[25]. In this experiment, we used two test collections. One is a subset of NACSIS-IR database which consists of abstracts presented at academic conferences sponsored by 24 Japanese academic societies¹, which used in the NTCIR Workshop[21]. This data set consists of 327,880 abstracts. The

¹<http://www.rd.nacsis.ac.jp/~ntcdm/index-en.html>

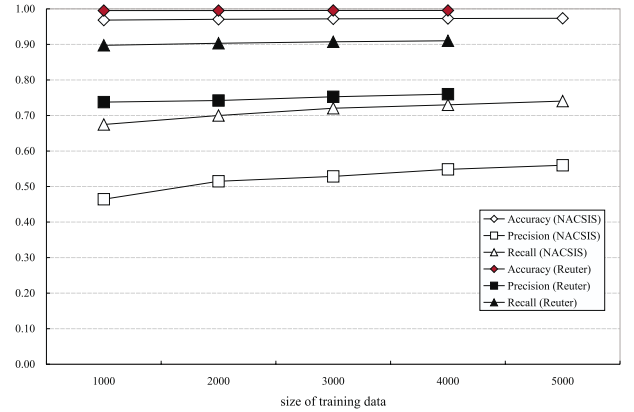


Figure 7: Results of Classification by SVM

other is the Reuters-21578 data set² compiled by David Lewis and originally collected by the Carnegie group from the Reuters newswire in 1987, which is one of standard test collections for text categorization community.

In order to construct feature vectors of articles, 1,000 words are selected as features based on the information gain criterion. Each document is represented with a term frequency vector of the selected 1,000 words weighted by tf-idf. We prepared 30 sets of training data for each size of training data ranging from 400 to 10,000 and a test data containing 10,000 abstracts. We applied SVM with Gaussian radial basis function kernel to those data sets. SVM^{light}³ was used to obtain SVM classifies.

The results of the experiment are shown in Figure 7. The average accuracy of SVM classifiers for each training data is over 95%.

We use a Sun UltraSPARC-II of 360MHz with 768MB of RAM, running Solaris 2.6. Learning of 5,000 training data took up to 10 CPU seconds. We suppose that this duration must be longer as reaction time at the interactive system but acceptable as a background process. We, therefore, should implement category customization process not as real-time process but as background one. We believe that SVM is practicable in aspects of accuracy and process time.

²<http://www.research.att.com/~lewis/reuters21578.html>

³http://www-ai.informatik.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html

6. CONCLUSIONS

This paper introduced an IR interface with user feedback which visualizes categories of documents and can be customized through user feedback with intuitive examinations.

The features of the proposed approach are shown as follows:

- Visualization of document categories helps users to recognize information space inclusively.
- The system can be customized in user feedback process with intuitive examinations.
- Our approach doesn't make its user examine documents which are retrieved with his/her initial query, but lets the user modify the category of some documents of his/her familiar domain.
- Clustering which is unsupervised learning method is used only for initial categorization of documents, and that reduces the necessity of computational resources for global analysis.
- SVM is used only for categorization of documents and user feedback.
- In addition to original features, user-depend attributes are used for representation of users' implicit cognitive aspects.

HCI research which should be actually effective for the users need rich real experiments. Therefore, we should implement a usable prototype system and make a great deal of experiments.

7. ACKNOWLEDGMENTS

Thanks are due to Akiko Aizawa of National Institute of Informatics for providing the test collection to us.

This research is a part of the research project "Studies on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources" (JSPS-RFTF96P00602) granted by JSPS (The Japan Society for the Promotion of Science).

REFERENCES

- [1] K. Aihara and K. Hori. Enhancing creativity through reorganizing mental space concealed in a research notes stack. *Knowledge-Based Systems*, 11(7-8):469–478, 1998.
- [2] K. Aihara, H. Koichi, and S. Ohsuga. Aiding the process of building knowledge out of chunks of information. *Journal of Japanese Society for Artificial Intelligence*, 11(3):432–439, 1996. in Japanese.
- [3] K. Aihara and A. Takasu. Domain visualization based on authorized documents. In *Proceedings of the fourth International Conference on Information Systems, Analysis and Synthesis*, volume 2, pages 391–398, Orlando, Florida, 1998.
- [4] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the university of michigan digital library. In *AAAI Spring Symposium on Information Gathering in Heterogeneous, Distributed Environments*, pages 19–24, 1995.
- [5] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.
- [6] T. Catarci and I. F. Cruz. Information visualization. *SIGMOD Record*, 25(4), 1996.
- [7] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR '92*, pages 318–329, 1992.
- [8] J. J. Daniels and E. L. Rissland. A case-based approach to intelligent information retrieval. In *ACM SIGIR '95*, pages 238–245, 1995.
- [9] S. Dao and B. Perry. Applying a data miner to heterogeneous schema integration. In *Proceedings of First International Conference on Knowledge Discovery and Data Mining*, pages 63–68, 1995.
- [10] S. Deerwester, S. T. Adumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
- [11] D. Dubin. Document analysis for visualization. In *ACM SIGIR '95*, pages 199–204, 1995.
- [12] U. Fayyad and R. Uthurusamy. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11), 1996.
- [13] R. S. Flounoy, R. Ginstrom, K. Imai, S. Kaufmann, G. Kikui, S. Peters, H. Schütze, and Y. Takayama. Personalization and users' semantic expectations. In *Query Input and User Expectations, Proceedings of SIGIR Workshop*, pages 31–35, 1998.
- [14] J. Hammer, H. Garcia-Molina, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. Information translation, mediation, and mosaic-based browsing in the tsimmis system. In *Exhibits Program of the Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 483–487, 1995.
- [15] M. Z. Hasan, A. O. Mendelzon, and D. Vista. Applying database visualization to the world wide web. *SIGMOD Record*, 25(4):45–49, 1996.
- [16] M. Hemmje, C. Kunkel, and A. Willett. Lyberworld – a visualization user interface supporting fulltext retrieval. In *ACM SIGIR '94*, pages 249–259, 1994.
- [17] D. A. Hull, J. O. Pedersen, and H. Shütze. Method combination for document filtering. In *Proceedings of SIGIR*, pages 279–298, 1996.
- [18] M. Iwayama and T. Tokunaga. Cluster-based text categorization: A comparison of category search strategies. In *ACM SIGIR '95*, pages 273–280, 1995.
- [19] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Technical report, Computer Science Department, University of Dortmund, 1997.
- [20] J. Mostafa, S. Mukhopadhyay, W. Lam, and M. Palakal. A multilevel approach to intelligent information filtering: Model, system, and evaluation. *ACM Transactions on Information Systems*, 15(4):368–399, 1997.
- [21] National Center for Science Information Systems. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.
- [22] M. Nodine, B. Perry, and A. Unruh. Experience with the infoleuth agent architecture. In *Proceedings of AAAI-98 Workshop on Software Tools for Developing Agents*, 1998.
- [23] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining*, pages 252–257, 1995.
- [24] M. Sugimoto, N. Katayama, and A. Takasu. COSPEX: A system for constructing private digital libraries. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, 1997.
- [25] A. Takasu and K. Aihara. Variance based classifier comparison in text categorization. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.
- [26] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(4):577–597, 1988.
- [27] Y. Yang and X. Lue. A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.