

Information Filtering via Fuzzy Hierarchical Induction*

Costin Barbu

Electrical Engineering
and Computer Science
Department
Tulane University
New Orleans, LA,
U.S.A.
barbu@eecs.tulane.edu

Rafal A. Angryk

Electrical Engineering
and Computer Science
Department
Tulane University
New Orleans, LA,
U.S.A.
angryk@eecs.tulane.edu

Fred E. Petry

Electrical Engineering
and Computer Science
Department
Tulane University
New Orleans, LA,
U.S.A.
fep@eecs.tulane.edu

Marin Simina

Electrical Engineering
and Computer Science
Department
Tulane University
New Orleans, LA,
U.S.A.
simina@eecs.tulane.edu

Abstract - *An adaptive algorithm is proposed in this work for learning the user profile based on his initial profile and on queries' interpretation using fuzzy concept hierarchies. The dynamics of the user profile is modeled by employing a new concept, Time - Words Vector Hyperspace. The preliminary results from applying this new approach are promising. Future plans and recommendations for further expanding are provided.*

Keywords: Information filtering, fuzzy induction, user profile.

1 Introduction

The list of documents returned by a Web search engine for a given query could easily exceed the thousands. In this work we investigate the role of the user profile within the frame of the current search context for filtering out the irrelevant documents.

An adaptive algorithm for learning the changes in user interests is presented in this paper. The vector hyperspace model introduced by Salton and McGill [13] has become a popular representation of the documents (and queries) in information retrieval. The expanded version of Salton's vector space model, introduced by Barbu and Simina [3] is used in this work for its effectiveness in computing the dynamics of the user profile. Comparing to the classical vector space model, this recent model, time-words vector hyperspace, has an additional temporal dimension. It considers the user current interests and their decay in time, if interests change. The coordinates of the documents and queries vectors are calculated using the traditional TF-IDF technique. Only the queries have a temporal dimension (current interest weight) which is set to a preset positive initial value that decays in time, suggesting that some specific user interests could decrease as time goes on. Hypernyms extracted by WordNet¹ and fuzzy concept

hierarchies are used in order to find the user's categories of interest.

The rest of the paper is organized as follows. Section 2 presents related work and its limitations. Section 3 introduces the user profile modeling and learning algorithm. Sections 4 and 5 discuss experimental results. Our conclusions and future work are presented in the final section.

2 Related work

Hierarchies for fuzzy summarization of database records were approached in late nineties by different groups of researchers. Lee and Kim [7] used ISA hierarchies, from the area of data modeling, to generalize database records to more abstract concepts. Lee [8] applied fuzzy generalization hierarchies to mine generalized fuzzy quantitative association rules. Cubero et al. [6] introduced fuzzy gradual rules for data summarization.

Many research surveys document various systems designed to provide user personalization while he browses the WWW. INFOS is a system that learns automatically by adapting its user model [10]. The user interests in different domains are represented by feature vectors. Keyword-based and knowledge-based techniques are employed for feature vector manipulation. The accuracy over keyword approach is improved by the hybrid approach. It also supports domain knowledge and retain the system's scalability.

Balabanovic [4] proposes an adaptive agent for Web browsing. The user profile is represented by a single feature vector weighted using the TF-IDF technique. The vector weight is increased or decreased based on the explicit positive or negative user's feedback.

* 0-7803-8566-7/04/\$20.00 © 2004 IEEE.

¹ WordNet is an on-line lexical reference system available at: <http://www.cogsci.princeton.edu/cgi-bin/webwn1.7.1>

Other authors explored genetic algorithms to learn user interests by incremental relevance feedback in NewT [14], and Amalthea [11]. Widyantoro developed Alipes [17], an intelligent agent that learns user's interests and provides personalized news articles retrieved from the Internet.

Although most of the mentioned works deal with learning user's profile, they do not emphasize on the adaptation of their systems to the changing of the user interests, except for the work of Widyantoro. Nevertheless the dynamics and the rate of change of the user interests were not addressed in previous work.

These problems have been initially addressed by the adaptive algorithm for learning the changes in user interests introduced by Barbu and Simina [3] who were using a heuristic approach. In contrast to that work, the current approach has a more complete foundation derived from the fuzzy induction theory.

3 User profile modeling and learning

3.1 Fuzzy concept hierarchy for extraction of generalized user interests

In this section we provide a simple technique, derived from the data-mining idea of Attribute-Oriented Fuzzy Induction [1], which allows us to extract the generalized sense (user's category of interest) from the query keywords. This approach is based on the generalization of keywords by employing background knowledge about these words' senses, coming from the WordNet.

WordNet is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University. Its design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet also provides various senses for a given word and their corresponding hypernyms.

Multiple senses of the query keywords can be modeled with the Fuzzy Concept Hierarchy. We define here a Fuzzy Concept Hierarchy *FCH* as an ordered pair (C, L) , where C is a set of concepts describing a particular word from the user's query and L is a set of links between these concepts. Each concept c has its unique name (label) and is a separate node in the hierarchy. A link l in the fuzzy concept hierarchy is a directed arc (edge) between two nodes with a certain weight assigned to it. Each link l , reflecting the operation of generalization (i.e. extraction of the abstract sense), can be uniquely characterized as a triple (s, t, μ_{st}) , where s and t are nodes and endpoints of l ,

and μ_{st} represents the share with which the single lower-level concept s belongs to the more general descriptor t .

The concept s is called a *source* of l and always placed on the lower level of abstraction in the generalization hierarchy; the second endpoint t is called commonly a *target* of l and we will refer to it as to a *direct abstract* of s [2].

Now we are able to build a Fuzzy Generalization Hierarchy for each of the query's keywords. First, using WordNet, we extract all hypernym chains for all senses corresponding to the given query expression. A hypernym chain of the particular word is actually a tree-like structure, which includes all abstract descriptors of the original word (it reflects ontological relation called "kind-of" dependency). For instance, for a word *burgundy* WordNet provided us with the following hypernym chains:

Sense 1:

- => French region
 - => geographical area
 - => region
 - => location
 - => entity

Sense 2:

- => wine
 - => alcohol
 - => beverage
 - => food
 - => substance
 - => entity
 - => liquid
 - => fluid
 - => substance
 - => entity
 - => drug of abuse
 - => drug
 - => agent
 - => causal agent
 - => entity

Sense 3:

- => dark red
 - => red
 - => chromatic color
 - => color
 - => visual property
 - => property
 - => attribute
 - => abstraction

Obviously these senses can be represented in the form of FCH, which can reflect the relations that have *many-to-many* character, where a single concept s can have more than one direct abstract t and s can belong to the higher-level concept t to a certain extent, expressed by

$\mu_{st}, \mu_{st} \in [0,1]$. We build such hierarchy with a bottom-up approach, where on the bottom of the hierarchy, which corresponds exactly to the lack of concept abstraction, we place the original user’s query expression.

Membership Degrees. Since word *burgundy* has three direct abstract descriptors: *French region*, *wine*, and *dark red* we split the memberships degrees evenly: $\mu_{burgundy-French\ region}=0.(3)$, $\mu_{burgundy-wine}=0.(3)$, and $\mu_{burgundy-dark\ red}=0.(3)$. The even, linear distribution of memberships reflects the situation when we do not have preferences concerning any of the senses extracted from WordNet. Each of the presented relations can be also described as a triplet, e.g: (*burgundy*, *French region*, 0.3), (*burgundy*, *wine*, 0.3), and (*burgundy*, *dark red*, 0.3). In the next step of generalization, according to the extracted hypernym chain, the descriptor *French region* has only one direct abstract (i.e. *geographical area*). In order to preserve completeness of the fuzzy model, we assign exactly the same weight to the $\mu_{French\ region-geographical\ area}$ as of the link incoming to concept *French region*: $\mu_{French\ region-geographical\ area}=0.(3)$ since the preceding membership value in the analyzed generalization path (i.e. $\mu_{burgundy-French\ region}$) is 0.(3). If branching occurs, as it does after the concept *alcohol*, we have to evenly split the preceding membership value in order to preserve completeness: $\mu_{alcohol-drug\ abuse}=0.1(6)$, and $\mu_{alcohol-beverage}=0.1(6)$, as shown in Figure 1.

Formalizing the above-mentioned characterization of the proposed membership function, we have to preserve the following properties in order to have it consistently reflecting the multi-sense generalization of a single keyword from the query:

(1) If we denote the original query keyword by c^0 (a single concept at the 0-abstraction level, i.e. placed at the bottom of the FCH) and its all direct abstracts by a set at the first level of abstraction: $C^1=\{c_1^1, c_2^1, \dots\}$; then we have to preserve the following condition in order to keep the model’s consistency at the first step of generalization:

$$\sum_{i=1}^{|C^1|} \mu_{c^0 c_i^1} = 1.0 \quad (1)$$

In other words, the sum of weights assigned to the links outgoing from the original keyword has to be equal to the unity.

(2) If we denote a single node in the FCH as $c_p^k \in C^k$, where k symbolizes the level of abstraction (we assumed upward numeration of abstraction levels in the FCH), and C^k symbolizes a set of all abstract concepts at the given k -abstraction level; then we have to assure the below condition in order to maintain the completeness:

$$\sum_{i=1}^{|C^{k-1}|} \mu_{c_i^{k-1} c_p^k} = \sum_{j=1}^{|C^{k+1}|} \mu_{c_p^k c_j^{k+1}} \quad \forall k > 0, \forall p \leq |C^k| \quad (2)$$

In other words, the sum of weights from all links incoming to each node in the FCH has to be equal to the sum of weights in all outgoing edges.

Rationale standing behind this approach is quite obvious: when the sum of all memberships derived during generalization of the concept *burgundy* is equal to 1.0 at each level of the concept hierarchy (at each generalization step), the completeness of the model is preserved. It guarantees the word *burgundy* to be constantly represented as a single concept at each level of abstraction. The complete fuzzy generalization hierarchy of the concept *burgundy* is presented in the Figure 1.

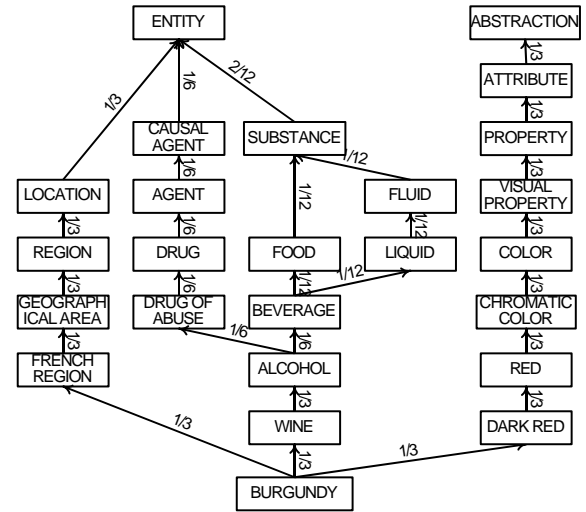


Figure 1. Fuzzy Generalization Hierarchy for the concept *burgundy* extracted from the hypernym chains.

The user category of interest is found by intersecting query’s keywords generalization models. For each of the generalization paths ending at the same abstract concept we calculate the Average Generalization Path Value (AGPV) as the average of the paths weights’ product, and choose the abstract with the largest value as the category of interest.

An example of finding the user category of interest given the *Query = {Burgundy, Cherry}* is illustrated in Figure 4. Assume the user recent profile is *Recent Profile = {physics, car, art}*. The Average Generalization Path Value is computed for both abstract concepts: *Food* and *Red*. $AGPV_{Food}(Query)$ is obtained by taking into consideration the mean average between path weights’ product on the chains: *Cherry – Food* and *Burgundy – Food*. $AGPV_{Red}(Query)$ is calculated employing the path weights’ product on the chains *Cherry – Red* and

Burgundy - Red. The FCH of the two candidate abstracts are intersected with the FCH of the previous category from the recent profile, *Art*. Since there are no intersections at the lower level of abstraction the AGPV of the candidate abstracts are compared. The abstract *Red* was selected as the category of interest since $AGPV_Red(Query) > AGPV_Food(Query)$.

Algebraic mean, employed to calculate the final AGPV, appropriately reflects our intentions. Due to applying the average (AVG) operator, the query's poly-semantic keywords have less influence on the extraction of the generalization concept (user's category of interest) from the query. On the other hand, applying the *product operator* for the AGPV computation increases the significance of generalized concept occurring at the lower level of abstraction over the common descriptors placed at the top of the hierarchy.

3.2 Learning the user profile

A new algorithm is introduced in this section for dynamic learning of user interests based on his initial profile and on queries semantic analysis using fuzzy concept hierarchies.

Documents and queries are represented as vectors in the time-words vector hyperspace and their weights are computed using the classical TF-IDF technique. In contrast with the document feature vectors, the queries feature vectors have an additional temporal dimension (current interest weight) set to a preset positive initial value that decays in time, suggesting that some specific user interests could decrease as time goes on. On the other hand the user interest can be held constant or raised if his search is related to already existing categories in his current profile.

The rate of interest change α_i computed as the cosine similarity between two sequential query feature vectors Q_i and Q_{i-1} proved to be an effective measure of user's profile dynamics [3]. The categories of interest are calculated based on a novel approach using fuzzy concept hierarchies and the WordNet ontology.

Two queues with similar structures are employed by our model to keep track of both user's Recent and Long-Term Profiles. The user's Long Term Profile queue has a larger capacity than the Recent Profile queue. The categories of interest extracted by the algorithm are stored in the Recent Profile queue (as illustrated in Figure 3) as long as the Current Interest Weight W_i is positive. The corresponding triplet (C_i, W_i, α_i) is transferred to the rear of the Long-Term Profile queue when W_i gets negative values. A similar action occurs if the Recent Profile queue is at its capacity. On the other hand, the triplet (C_i, W_i, α_i) is removed from the front of the Long Term Profile queue

when it reaches its capacity. The Current Interest Weight W_i has been considered to decrease linearly during the Recent Profile time and exponentially within the Long Term Profile period of time.

User Profile Learning Algorithm. Given a preliminary user profile, at most L domains of interest for the Long Term Profile, R domains of interest for the Recent Profile, and M the preset number of elements of a vector, the algorithm *LearnUserProfile* for learning the user profile from queries within the current search context is defined as follows.

```

Learn_User_Profile( $Q_i$ ) => updated profile  $P$ 
1. Call subroutine Extract_Category( $Q_i$ )
2. Compute the Rate of Interest Change  $\alpha_i$  between  $Q_i$  and  $Q_{i-1}$ 
3. Insert Category  $C_i$ , to the Recent Profile together with the Current Interest Weight  $W_i$  (preset to a positive initial value  $W$ ) and with the Rate of Interest Change  $\alpha_i$ 
4. If Rate of Interest Change  $\alpha_i > \alpha_{threshold}$  ( $= 0.6$ ) then increase the Current Interest Weight of Category  $C_i$  by a positive value  $\Delta W : W_i = W + \Delta W$ 
5. Sort the triplets  $(C_i, W_i, \alpha_i)$  from the Recent Profile in ascending order of the Current Interest Weight  $W_i$ .
6. Decrease linearly all  $W_i$  from Recent Profile by a temporal decay factor  $\theta_1$ 
7. If  $W_i < 0$  then move  $(C_i, W_i, \alpha_i)$  to the Long Term Profile
8. Decrease exponentially all  $W_i$  from Long Term Profile by a temporal decay factor  $\theta_2$ 
9. Return Updated User Profile.

```

```

Extract_Category( $Q_i$ ) => Category  $C_i$ 
1. For each query  $Q_i = \{t_{1i}, t_{2i}, t_{3i}, \dots, t_{ki}\}$ , where  $k = 1 \dots M$  and  $t_{ki}$  are the keywords of  $Q_i$ 
2. Intersect the FCH of keywords  $t_{ki}$  to find abstract terms Term1, Term2, ...
3. IF there exist Term1, Term2
4. CALL Extract_Category(Term1, Term2,  $C_{i-1}$ )
5. IF FCH of Term1, Term2,  $C_{i-1}$  intersect in the GENERAL ABSTRACTION (i.e. Entity, Phenomenon)
6. Select  $C_i =$  Term that corresponds to  $\max(AGPV\_Term_1(Q_i), AGPV\_Term_2(Q_i))$ 
7. ELSE
8. IF  $\max(AGPV(Term_1, C_{i-1})) > \max(AGPV(Term_2, C_{i-1}))$ 
9. Select  $C_i = Term_1$ 
10. ELSE
11. Select  $C_i = Term_2$ 
12. END
13. END
14. END

```

Figure 2. Algorithm *Learn_User_Profile* and subroutine *Extract_Category*

Recent Profile

dance	music	food	art	Categories
100	85	75	60	Current Interest Weight
0.65	0.35	0.30	0.25	Rate of Interest Change

Figure 3. User Recent Profile representation

4 Experimental results

Let's consider the user's query being $Query = \{burgundy, cherry\}$ and his recent profile is $Recent Profile = \{Physics, Car, Blue\}$. WordNet generates the hypernym chains for the concept cherry as illustrated in Figure 3.

$Sense 1 :=> wood => plant material => material => substance => entity$
$Sense 2 :=> fruit tree => tree => woody plant => plant => organism => living thing => object => entity$
$Sense 3 :=> edible fruit => produce => food => solid => substance => entity$
$\Rightarrow drupe => fruit => plant organ => plant part => natural object => object => entity$
$Sense 4 :=> red => chromatic color => color => visual property => property => attribute => abstraction$

Figure 4. Hypernym chains for the concept cherry

According to our approach we can extract the user category of interest by intersecting both generalization models and finding the first (the lowest) occurrence of the common abstraction as shown in Figure 5. For each of the generalization paths ending at the same abstract concept we calculate the Average Generalization Path Value, and choose the abstract with the strongest path (largest value) as the category of interest and added to user's recent profile.

(1) Sense Red

Path from *Burgundy* to Red: $\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{9} = 0.11$

Path from *Cherry* to Red: $\frac{1}{4} = 0.25$

So the *AVG* of the paths leading to the concept Red is $AGPV_Red = (0.11+0.25)/2=0.36/2=0.18$

(2) Sense Food

Path from *burgundy* to Food: $\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{648} = 0.0015$

Path from *cherry* to Food: $\frac{1}{8} \cdot \frac{1}{8} \cdot \frac{1}{8} = \frac{1}{24} = 0.0417$

So the *AVG* of the paths leading to the concept Food is $AGPV_Food = (0.0015+0.0417)/2=0.0432/2=0.0216$. Therefore we extract the lowest common abstraction as Red (since it has the strongest path), and update user's profile with it: Recent Profile = {Physics, Car, Art, Red};

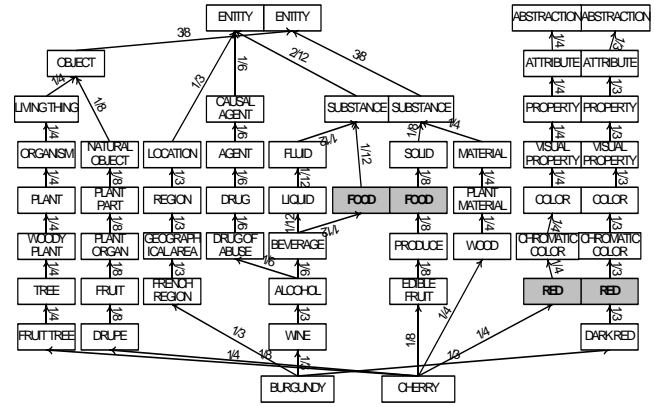


Figure 5. Intersection of Fuzzy Generalization Hierarchies for Query keywords: burgundy and cherry.

5 Information filtering based on the profile dynamics

Here are some experimental results we have gathered by testing various information filtering methods, based on the user profile dynamics, as shown in Table 1. The search efficiency increases if the retrieved documents are filtered according to the Relevance Score. This score can be determined as the cosine similarity between the User Profile feature vector and the feature vectors of the documents retrieved by a classical search engine (i.e. Google).

	No Profile Filtering	Total Profile Filtering	Recent Profile Filtering
Total Number Documents Retrieved	40,174,000	312	15,432
Accuracy in Top 10 Documents Retrieved	39.66 %	50.75 %	75.33%
Accuracy in Top 20 Documents Retrieved	29.33 %	35 %	67.66 %

Table 1. Experimental results (on average) for various filtering methods on a set of queries

The search performance is significantly improved per our preliminary results by employing the contextual relevant information and by considering the rate of user's

interest change. The average accuracy (relevance rate) using the Recent Profile Filtering method outperforms the other filtering methods accuracies for both the top 10 and top 20 documents retrieved cases, as shown in Table 1.

6 Conclusions

We have presented an adaptive algorithm for learning the changes in user interests based on his initial profile and on user's queries. We introduced a new approach based on fuzzy concept hierarchies for extracting the user profile from his queries. The information filtering effectiveness is significantly improved by using the User Recent Profile as opposed to existing approaches that consider the total profile [4], [5], [17]. Our work can be extended in order to accommodate queries with brand names by integrating the previous knowledge available in WordNet with a more specialized ontology.

References

- [1] R. Angryk, and F. Petry, "Data Mining Fuzzy Databases Using Attribute-Oriented Generalization", Proc. of the Third IEEE International Conference on Data Mining, Melbourne, FL, November 2003.
- [2] R. Angryk, and F. Petry, "Consistent Fuzzy Concept Hierarchies for Attribute Generalization", Proc. of the IASTED International Conference on Information and Knowledge Sharing, Scottsdale, AZ, November 2003.
- [3] C. Barbu, and M. Simina, "Information Filtering Using the Dynamics of the User Profile", Proc. of the 16th International Florida Artificial Intelligence Research Society Conference, St. Augustine, FL, pp. 245-249, May 2003.
- [4] M. Balabanovic, "An Adaptive Web Page Recommendation Service", Proc. of the First International Conference on Autonomous Agents", New York, pp. 378 – 385, 1997.
- [5] L. Chen, and K. Sycara, "WebMate: Personal Agent for Browsing and Searching". Proc. of the Second International Conference on Autonomous Agents, New York, pp. 132-139, 1998.
- [6] J. C. Cubero, J.M. Medina, O. Pons, and M.A. Vila, "Data Summarization in Relational Databases Through Fuzzy Dependencies", *Information Sciences*, Vol. 121, No.3-4, pp. 233-270, 1999.
- [7] D. H. Lee, and M. H. Kim, "Database Summarization Using Fuzzy ISA Hierarchies", *IEEE Transactions On Systems, Man, and Cybernetics* - part B, Vol. 27, No. 1, pp. 68-78, 1997.
- [8] M. Lee, "Mining Generalized Fuzzy Quantitative Association Rules with Fuzzy Generalization Hierarchies", Proc. of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, pp. 2977 – 2982, 2001.
- [9] M. McElligot, and H. Sorensen, "An Evolutionary Connectionist Approach to Personal Information Filtering", Proc. of the Fourth Irish Neural Network Conference, Dublin, Ireland, pp. 141-146, 1994.
- [10] K. J. Mock, "Hybrid-Hill-Climbing and Knowledge-based Techniques for Intelligent News Filtering", Proc. of the 13th National Conference on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence Conference, Menlo Park, CA, pp. 48-53, 1996.
- [11] A. Moukas, and G. Zacharia, "Evolving a Multiagent Information Filtering Solution in Amalthea", Proc. of the First International Conference on Autonomous Agents, New York, N.Y, pp. 394-403, 1997.
- [12] G. Raschia, L. Ughetto, and N. Mouaddib, "Data Summarization Using Extended Concept Hierarchies", Proc. of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, BC, Canada, pp. 2289 -2294, 2001.
- [13] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York. N. Y. 1993.
- [14] B. D. Sheth, "A Learning Approach to Personalized Information Filtering", M.S. diss., Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1993.
- [15] A. Tan, and C. Teo, "Learning User Profile for Personalized Information Dissemination", Proc. of 1998 International Joint Conference on Neural Networks, Anchorage, AK, pp. 183-188, 1998.
- [16] E. M. Voorhees, "Using WordNet to Disambiguate Word Senses for Text Retrieval", Proc. of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1993.
- [17] D. H. Widyantoro, J. Yin, M. S. El Nasr, L. Yang, A. Zacchi, and J. Yen, "Alipes: A Swift Messenger in Cyberspace", Proc. of the Spring Symposium on Intelligent Agents in Cyberspace, Palo Alto, CA, pp. 62-67, 1999.
- [18] E. Wiener, J. Pederson, and A. Weigend, "A Neural Network Approach to Topic Spotting", Proc. of the Fourth Annual Symposium on Document Analysis and

Information Retrieval, Las Vegas, NV, pp. 317-332,
1995.