Toward Automated Research Topics Discovery from WWW with KAROKA

David Ramamonjisoa Faculty of Software and Information Science, Iwate Prefectural University
david@soft.iwate-pu.ac.jp

Keywords: research topics, web mining, agents, association rules

Summary

In this paper, we present agents based tool to discover new research topics from the information available on the World Wide Web (WWW). Agents are using KAROKA (Keywords Association Rules Optimizer Knobots Advisers). KAROKA is a model of discovery in text database used in WWW. The WWW sources are converted to a highly structured collection of text. Then, KAROKA tries to extract topics, association rules, regularities and useful information in the collection of text. KAROKA techniques are described such as information retrieval similarity metrics for text, generation and pruning of keywords combination, and summary proposal of discovered information.

1. Introduction

When a user explores a new domain, attempting to summarize the essence of an area previously unknown to the user, it is called *information and knowledge discovery* [Crimmins et al. 99]. Imitating that activity in a machine is the main research of web mining and information discovery agents [Cooley et al. 97, Levy et al. 99, Chakrabarti et al. 99].

Information Discovery Agents (called also Web Mining Agents) are a set of an important kind of information seeking system trying to realize the previously mentioned tasks. The agents are built with the artificial intelligence techniques and information retrieval methodologies such as automated text categorization, machine learning, topics detection and tracking, clustering, and probabilistic models [Mitchell 96, Sebastiani 99, Miller et al. 99].

Current systems are focusing on the quality of the result according to the users relevance feedback and preferences. Some systems are using sophisticated algorithms to optimize the quality of extracted information and knowledge. These methods do not bring novelty, utility, and understandability to the results [Pazzani 00]. Intelligent Information Discovery Agents should have also some mechanisms to distinguish the irrelevant data and unexpected interesting results.

The domain we focus on is the discovery of Research Topics on "Data Mining and Knowledge Discovery and Their Applications." We try to determine: "What are possible and promising research topics according to the information on the Web in this research domain?"

We introduce *KAROKA (Keywords Association Rules Optimizer Knobots Advisers)*, a personalized system that pro-actively tries to discover information from various distributed sources and presents it to the user in the form of a digest. KAROKA is using a tool similar to "AltaVista Discovery" [Altavista 00] or CiteSeer [CiteSeer 00] for the exploration of World Wide Web.

In section 2, we describe our KAROKA model. Section 3 explains the association rule mining. Section 4 details the example of KAROKA use and experimental results in *"data mining trends and forecasts."* Section 5 summarizes the paper and describes our future work.

2. KAROKA Model and Architecture

KAROKA objectives are to design agents that can process user queries in domain specific research areas, collect WWW sources relevant to the queries as a research corpus, extract keywords and rules from the retrieved sources, infer or induce to determine possible new research topics in the domain, and present results as *'list of possible new research topics'* in the

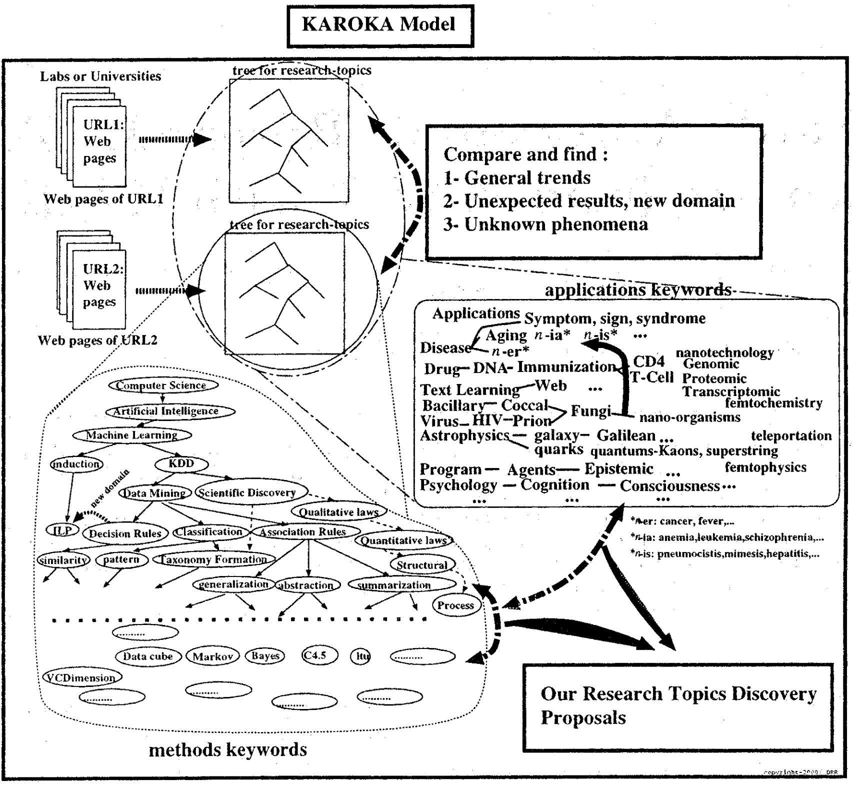

Fig. 1 KAROKA model and analysis : User focuses to the research topics in found web pages by a search engine according to a query in a specific domain (in this figure, the domain is in *Computer Science/Artificial Intellicenge/Machine Learning/Knowledge Discovery and Data mining*). Documents are retrieved from *URL* and are structured to research topics tree. *Research topics* are composed of one or several keywords. Keywords are classified into *method keywords* or *application keywords*. According to our strategies and algorithms, new research topics are derived

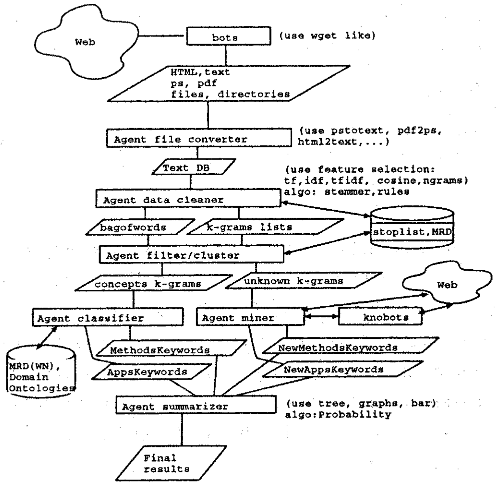

Fig. 2 KAROKA agent architecture

domain. KAROKA (see figure (Figure 1)) uses keywords extracted from technical and project presentation articles available in the World-Wide-Web documents.

KAROKA does not search the WWW itself but instead launches multiple agents that utilize existing indexing engines and perform a "meta-search" in order to collect and discover information that is broadly of interest to the user. Then the system further analyzes the retrieved documents in building keywords database and structured tree such as indexes to the research topics domain. KAROKA knobots monitor frequently changing information resources and update the database.

According to the research topic area, it generates and combines randomly keywords of research subtopics. Strategies and constraints for the new topics selection are based on classification techniques, association rules, and verification on the WWW.

Our agents architecture is depicted in the figure (Figure 2). Each agent's input and output is detailed with the internal or external knowledge or data bases. Agents interface is written on scripting language Perl. Internal processes and algorithms are coded in conventional programming language such as C or Java. Each task is described in details below.
1) Some research topics are already indexed and categorized hierarchically in the search engines via the Web. Those research topics are general and common. For example, the hierarchy of the research topics on Data Mining and Knowledge Discovery has many subtopics as represented in table (Table 1).

The first task concerns the preparation of the corpus.
After the user entered a query from an interface, the system uses search engine to find general 'research topics' to the research domain he/she is interested. The system selects some URLs from the search results to start the discovery of new research topics. For each URL pages, there is a list of contents. The system focuses on the content *'research topics or research areas'* if it exists. The system retrieves all documents in found section and stores in his computer. If the document is in HTML format, then document is transformed to structured text, the headers are treated as a special type of keyword.
2)The second task is the keyword extraction.
One obvious methods to extract keywords is to find the keywords as the authors defined in the document. Some documents such as articles or technical papers, research reports contain explicitly the keywords. These *explicit keywords* are treated with priority. By experience, we state that they are potential *research topics.*
Documents without explicit keywords are processed with the document representation (bag of words) [Salton 89]. By using our categorization model, we can extract the important words in the bag of words. Some rules are created during the extraction. Rules concern to separate the words with low frequency and high weight.
The agent filtering and clustering allow to clean the data by using feature selection and to classify the two kinds of topics as known or unknown.
3)The classification and mining tasks.
The classifier as its name classifies the keywords as "method keywords" and "application keywords." The "method keywords" are related to the research topics and their sub-topics generally already known in the research community. The "application keywords" are related to the other domains that the methods are applied. Domain ontologies are used according to their availability.
The mining task :
First keywords selection consists with the elimination of high frequency keywords. They are too common for the topics. Remaining keywords from the first selection are combined two-by-two to obtain the trends of the research within these research topic keywords. We then check with the knobot adviser module the relevance of these generated combination of keywords

Table 1 Categorization of research topics

```
Computer Science>Artificial Intelligence>Machine Learning
Machine Learning>Knowledge Discovery>...
                 Data Mining>Application>...
                             Classification>...
                             Feature Selection>...
Classification>Decision Rules - Winnow - TFIDF - Naive Bayes - ...
...
```

in the Web.

According to the relevance of the keywords in the Web, a second selection is necessary to eliminate again the high frequency keywords. The same method as in first selection is used. At this stage, we observed the existence of research topics classes. The high frequency keywords class is belong to the known research topics. Low frequency keywords class may be new research topics or irrelevant keywords. However, this criteria does not have effect to the application keywords.

We applied the first selection method to the *application keywords.* We added the *application keywords* to two combined keywords result of the previous second keywords selection method. We then generated trends of research topics based on three keywords as two method keywords and one application keyword. A final check with the Web is realized to get the new research topics proposals.

An association rules module is used to link the classified topics to the new topics and estimate the maximum relevance. The criteria for the relevance are the occurrence and the rank of the related URL given by the search engine such as first 10 or 100 matches.

The knobot then checks the URL to confirm the relevancy by collecting possible technical papers or reports.

3. Keywords Association Rules

Basically, association rule mining searches for interesting relationships among items in a given data set. We adopted the notations proposed by the inventors of fast algorithm for association rules mining[Agrawal et al. 94].

Let $J=\{i_1,i_2,...,i_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transaction where each transaction T is a set of items such that $T \subseteq J$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset J$, $B \subset J$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with **support** s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has **conifidence** c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(A|B)$. That is,

$support(A \Rightarrow B) = P(A \cup B)$

$confidence(A \Rightarrow B) = P(A|B)$

Rules that satisfy both a minimum support threshold (*min_sup*) and a minimum confidence threshold (*min_conf*) are called **strong.**

A set of items is referred to as an **itemset**. An itemset that contains k items is a **k-itemset**. The occurrence frequency of an itemset is the number of transactions that contain the itemset. An itemset satisfies **minimum support** if the occurrence frequency of the itemset is greater than or equal to the product of *min_sup* and the total number of transactions in D. If an itemset satisfies minimum support, then it is a **frequent** itemset.

Association rule mining is a two-step process:

(1) **Find all frequent itemsets**: each of these itemsets will occur at least as frequently as a predetermined minimum support count.

(2) **Generate strong association rules from the frequent itemsets**: these rules must satisfy minimum support and minimum confidence.

The keywords association rules are based with the apriori algorithm invented by Agrawal et al..

3·1 *Apriori algorithm*

Apriori algorithm is composed of a basic algorithm for finding frequent itemsets and a procedure for generating strong association rules from frequent itemsets.

§1 Support of Itemsets

Let X be the set of all item sets under consideration. The support of an itemset S is the percentage of those item sets in X which contain S. If $Y \subseteq X$ is the set of all item sets such that $\forall T \in Y$: $S \subseteq T$,

then:
$support(S) = |Y|/|X| * 100$

§ 2 Confidence of Association Rule

The confidence of a rule $R =$ 'A and $B \Rightarrow C'$ is the support of the set of all items that appear in the rule divided by the support of the antecedent of the rule, i.e.:
$confidence(R) = support(\{A, B, C\})/support(\{A, B\}) * 100$

4. Experiments

We first evaluated the association rule mining by using a program called **Apriori** developed by a Researcher in Germany[Borgelt 96].
Our dataset is a collection of computer science articles and technical reports documents. The collection has 160 documents. Each document has explicit keywords and those keywords are easily extracted to be the itemset for the association rule. The document filename is the transaction ID.
After the extraction, 648 keywords (or items) has been retrieved.

Combination of the keywords with apriori program returned 34674 rules by setting the minimum support to 0.1% and minimum confidence to 0.1%.
We eliminated the evident rules by removing the rules with support greater than 93% or confidence greater than 93%. Those rules reflects the keywords relationship within one document.

Rules with a range of support and confidence (5% $<s<$25%, 25% $<c<$80%) are proposed as results (see table Table 2). Discerning research topics among these rules remains a difficult task. A graph of keywords is proposed to form the relationship among 648 keywords of the 160 documents.

The graph keywords relationships should guide to the human user a meaningful research topics.

5. Conclusions and Future works

In this paper, we presented a model for research topics discovery from the Information World Wide Web. The model is based on KAROKA system. KAROKA is a personalized tool using keywords association rules and knobots. With KAROKA, we have partially automated the discovery.
Our experiment results show the KAROKA system applied to discover new research topics on *Data Mining and Knowledge Discovery.* In computer science research topics, we found that algebraic notations have strong concepts. Until now, they are ignored and are not possible to compute yet in the bag of words. The equivalent words may exist but they are ambiguous (e.g. cosine, chi-square, probability P).

At the stage of the KAROKA program, the user must interpret the result given by KAROKA as a support for his/her research topics finding. Our work is very close to the research described in [Sanderson et al. 99] for the derivation of hierarchical concepts, however the mining from WWW is not included in their work. In the future, we are refining the KAROKA program to be more useful such as in TopCat [Clifton 1999].

◇ References ◇

[Crimmins et al. 99] Crimmins, F. et al. :"TetraFusion: Information Discovery on the Internet," IEEE Intelligent Systems Journal, pp.55-62, July-August, 1999.

[Cooley et al. 97] Cooley et al.: "Web Mining: Information and Pattern Discovery on the World Wide Web", in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

[Levy et al. 99] Levy, A.Y. and Weld, D.S. : "Intelligent Internet systems" Artificial Intelligence Journal, vol. 118, numbers 1-2, pp.1-14, April, 2000.

[Chakrabarti et al. 99] Chakrabarti, S. et al. : "Focused Crawling: A New Approach for Topic-Specific Resource Discovery." in Proceedings of 8th International WWW Conference, Elsevier Science, pp.545-562, 1999.

[Mitchell 96] Mitchell T. :"Machine learning", McGraw Hill, New York, 1996.

[Sebastiani 99] Sebastiani F. : "A Tutorial on Automated Text Categorisation", in Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence", pp.7–35, 1999.

[Miller et al. 99] Miller D. et al. : "A Hidden Markov Model Information Retrieval System," in Proceedings of the ACM Sigir'99, pp.214–221, 1999.

[Pazzani 00] Pazzani, M. :"Trends and Controversies: Knowledge discovery from data ?" IEEE Intelligent Systems Journal, pp.10–13, March-April, 2000.

[Altavista 00] http://discovery.altavista.com

[CiteSeer 00] http://www.researchindex.com

[Chakrabarti et al. 98] Chakrabarti, S. et al. : "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies." in Journal of Very Large Databases 7,3, pp163–178, 1998.

[Cavnar et al. 94] Cavnar, W and Trenkle, J. : "N-gram-based text categorization," in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp.161–175, 1994.

[Agrawal et al. 94] Agrawal, R. and Ramakrishnan, S. : "Fast Algorithms for Mining Association Rules," in Proceedings of the 20th Very Large Database Conference, Santiago, Morgan Kaufmann, pp478–499, 1994.

[Hearst 99] Hearst, M. : "Untangling Text Data Mining," in the Proceedings of ACL'99, June, 1999.

[Schwartz et al. 97] Schwartz R. et al. : "A Maximum Likelihood Model for Topic Classification of Broadcast News, " in Proceedings of Eurospeech Conference, 1997.

Table 2 Rules with apriori in computer network domain. The rule has the form $B \Leftarrow A$ (support%, confidence%)

```
...
collective communication <- wormhole routing  (8.7%, 71.4%)
multicast <- wormhole routing  (8.7%, 64.3%)
broadcast <- wormhole routing  (8.7%, 64.3%)
parallel computer architecture <- wormhole routing  (8.7%, 28.6%)
path-based routing <- wormhole routing  (8.7%, 35.7%)
interprocessor communication <- wormhole routing  (8.7%, 28.6%)
meshes <- wormhole routing  (8.7%, 28.6%)
multicast <- collective communication  (8.1%, 76.9%)
collective communication <- multicast  (6.8%, 90.9%)
broadcast <- collective communication  (8.1%, 76.9%)
parallel computer architecture <- collective communication  (8.1%, 38.5%)
path-based routing <- collective communication  (8.1%, 38.5%)
interprocessor communication <- collective communication  (8.1%, 30.8%)
broadcast <- multicast  (6.8%, 90.9%)
parallel computer architecture <- multicast  (6.8%, 36.4%)
path-based routing <- multicast  (6.8%, 36.4%)
interprocessor communication <- multicast  (6.8%, 27.3%)
meshes <- multicast  (6.8%, 27.3%)
cut-through routing <- multicast  (6.8%, 27.3%)
parallel computer architecture <- broadcast  (6.2%, 40.0%)
path-based routing <- broadcast  (6.2%, 40.0%)
interprocessor communication <- broadcast  (6.2%, 30.0%)
meshes <- broadcast  (6.2%, 30.0%)
cut-through routing <- broadcast  (6.2%, 30.0%)
...
```

[Schwartz et al. 01] Schwartz R. et al. : "Unsupervised Topics Discovery " in Proceedings of Workshop on Language Modeling and Information Retrieval, pp.72-77, 2001.

[Sanderson et al. 99] Sanderson M. and Croft B. : "Deriving concept hierarchy from text," in Proceedings of the ACM Sigir'99, pp.206-213, 1999.

[Lavrenko et al. 00] Lavrenko, V. et al. : "Mining of Concurrent Text and Time Series," in the Proceedings of KDD 2000 Conference, pp. 37-44, 2000.

[Salton 89] Salton, G. : "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer;", Addison-Wesley, 1989.

[Borgelt 96] C. Borgelt : "Apriori program", free software foundation, 2000.

[Clifton 1999] C. Clifton, R. Cooley: "TopCat: Data Mining for Topic Identification in a Text Corpus.", PKDD 1999, p.174-183, 1999.