Concept Mining for Indexing Medical Literature

Isabelle Bichindaritz, Sarada Akkineni

University of Washington, 1900 Commerce Street, Box 358426,
Tacoma, WA 98402, USA
ibichind@u.washington.edu

Abstract. This article addresses the task of mining concepts from biomedical literature to index and search through this documents base. This research takes place within the Telemakus project, which has for goal to support and facilitate the knowledge discovery process by providing retrieval, visual, and interaction tools to mine and map research findings from research literature in the field of aging. A concept mining component automating research findings extraction such as the one presented here, would permit Telemakus to be efficiently applied to other domains. The main principle that has been followed in this project has been to mine from the legends of the documents the research findings as relationships between concepts from the medical literature. The concept mining proceeds through stages of syntactic analysis, semantic analysis, relationships building, and ranking.

1 Introduction

As the number of scientific publications is increasing tremendously, it is becoming hard for researchers to find relevant information in a domain and keep up with the information flow. Researchers are foremost interested in collecting and analyzing the research findings presented in research literature. There is a need for an information system that can be used by researchers to extract domain knowledge in the form of research findings without trying to read the whole article. The idea pursued in this article is to extract knowledge from articles using mostly tables and figures legends, since this is where the main research findings are the most likely to be represented.

With the rapidly growing body of scientific knowledge and increasing overspecialization in specific domains, it is likely that the scientific work of one research group might also solve an important problem that arises in the work of another group. Yet the two groups might not be aware of the work of each other. However, important knowledge is recorded at least in textual form in bibliographic databases, such as Medline for the field of biomedicine. In the present context these large documents databases provide both an opportunity and a need for developing advanced methods and tools for computer supported knowledge discovery [2, 3].

The goal of literature-based discovery in general is to discover new and potentially meaningful relations between a given starting concept of interest and other concepts by mining bibliographic databases [5, 6, 7] such as Medline [13]. The idea of discovering new relations from a bibliographic database was introduced by [13, 14], who describes a data mining system that made seven medical discoveries that have been later published in relevant medical journals.

Telemakus is one such information system dedicated to the idea of literature-based information mining [1] for a core set of concepts in a medical domain. The motivating idea of Telemakus project is to aid researchers in the caloric restriction and aging domain to rapidly find the research findings and main concepts from research articles [8] in this domain. The research findings are represented by relationships between pairs of such concepts. The system can then be used to find, in particular, all the articles that study a particular relationship. For instance, a researcher can use this system to find all the articles which studied the relationship between caloric restriction and aging. The most helpful knowledge any researcher can gain while entering a new domain is the set of research findings so far, which is what can be more easily acquired using Telemakus system [4]. Telemakus project is intended to assist researchers by providing visual and interaction tools to visualize and navigate the research findings in their domain of research [4].

The system presented here proposes to automate the process of extracting the research findings from the literature by mining the concepts and relationships from documents and indexing these by the concepts learnt. An important feature of this system is that it mines for relationships between concepts, such as the relationhip between caloric restriction and aging, and not for isolated concepts. The next section presents the Telemakus project. The third section sets forth the system architecture, and the fourth section explains the concept mining process in detail through its different stages. It is followed by the results and a conclusion.

2 Telemakus Project

The goal of the Telemakus project at the University of Washington Medical School is to support and facilitate the knowledge discovery process by developing interaction tools to retrieve and visualize documents by their research findings. For that purpose, this system mines and maps research findings from research literature. Telemakus system proposes tools and a framework to create, maintain, and query a database of documents through their research findings. Telemakus is part of SAGE KE, the Science of Aging Knowledge Environment and an online resource for researchers in the field of aging [15].

The Telemakus system [4] consists of a set of domain documents (current focus is the biology of aging), a conceptual schema to represent the main components of each document, and a set of tools to query, visualize, maintain, and map the set of documents through their concepts and research findings [4]. The conceptual schema is composed of standard bibliographic information, information about the research process (age, sex, number of subjects, treatment regimen, and research criteria for research animals), and most importantly research findings derived from data tables and figures. The “Unified Medical Language System” (UMLS), a specialized knowledge source in biomedicine, provides standardized concepts for the creation of a controlled domain vocabulary. The UMLS provides a very powerful resource for controlled domain vocabulary. The UMLS provides a very powerful resource for rapidly creating a robust scientific thesaurus in support of precision searching. Further, the semantic type descriptors for each concept and semantic network may offer some interesting opportunities for intelligent searching and mapping of concepts representing research findings, and their relationships. At present, knowledge extraction resorts to systems with both manual and automated components. A key area of current work is to move towards automating the research concept identification process, through data mining [4].

Fig. 1 shows the architecture of Telemakus system. *Fetcher*, *Extractor*, and *Crosscheck* are the most important components in the system. *Fetcher* fetches documents from the bibliographic database and stores them in the domain-specific database (referred to as DSDB). The extraction process is performed by a domain expert who manually interprets each legend to extract relevant concepts and uses his or her knowledge of the domain to establish relationships.

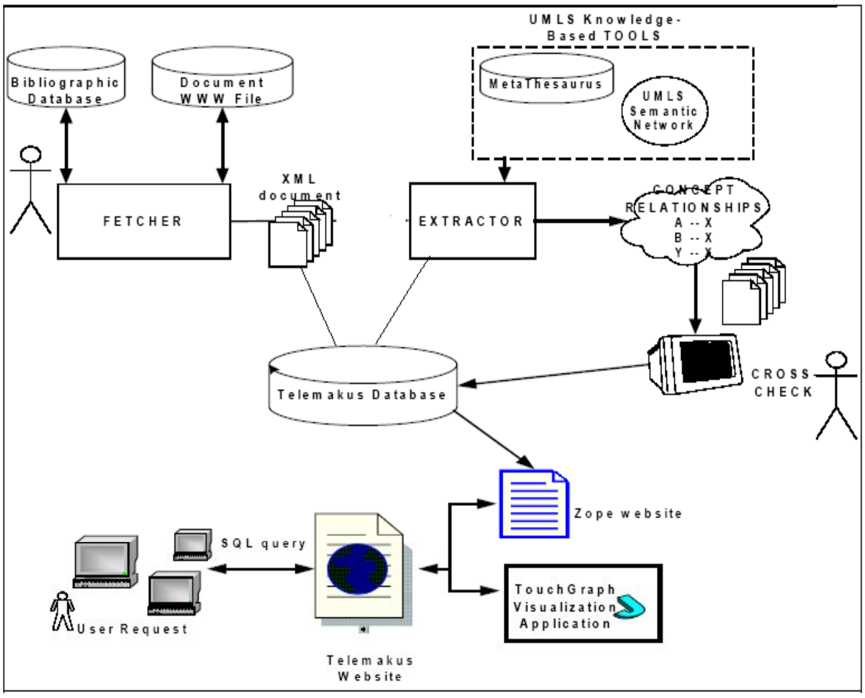


Fig. 1. Telemakus architecture

Crosscheck is a visualization tool used by researchers through the Web interface and by the domain expert for evaluation. The current Telemakus database has a Web interface, which can be accessed at **www.Telemakus.net**.

3 Architecture

The architecture of the systems (see Fig. 2) shows the datastores and components of the concept mining system. There are two knowledge bases involved, UMLS database, and DSDB database. Within DSDB, the domain specific thesaurus represents the standardized vocabulary of the "caloric restriction and aging" domain. The components of the system are the following:

1. *Data Access Component*, which extracts the document to mine from DSDB.
2. *Syntactic Analyzer*, which analyzes the syntax of a document by parsing its legends and extracting lexical information.
3. *Relationship Builder*, which takes the lexical information from above, locates a trigger phrase for a relationship from each legend, and forms from there a triple composed of two phrases and a trigger phrase. Each triple represents a relationship between two concepts.
4. *Relationship Selector*, which semantically analyzes each phrase in each triple by accessing the UMLS, and extracts from each phrase its main concepts.
5. *Ranker*, which ranks the relationships extracted.
6. *Evaluator*, which evaluates the result of the concept mining process for each document. It accesses DSDB through Data Access Component in order to get the results of the manual mining process, and compare them with the automated ones. It produces precision and recall ratios as the reference in evaluating the system.

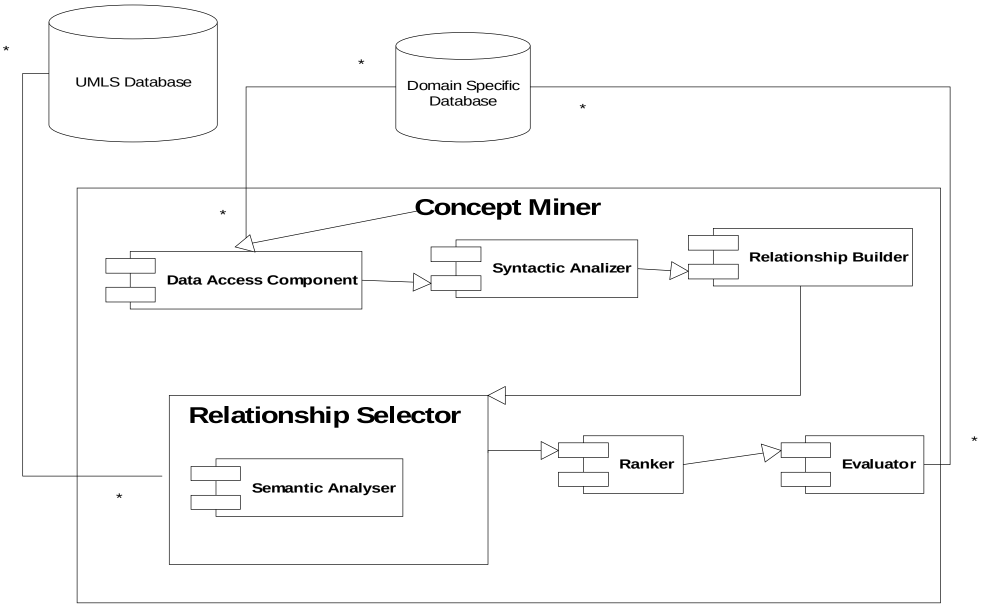
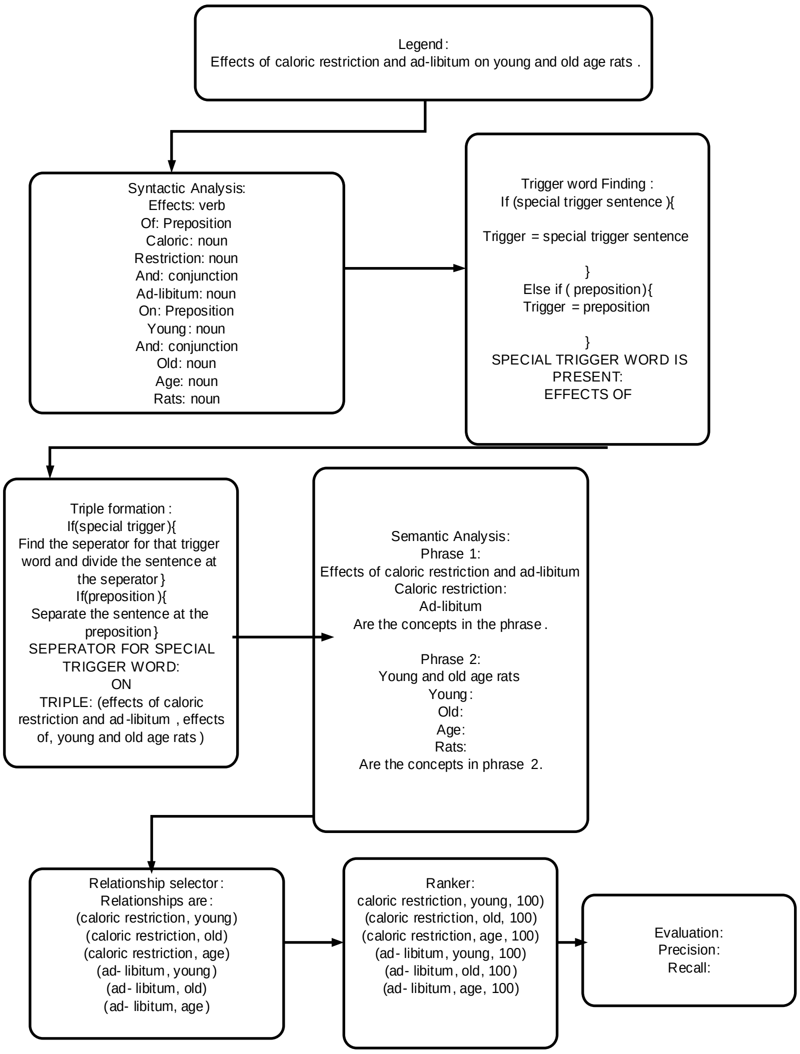


Fig. 2. Concept Miner architecture

4 Concept Mining Process

Concept mining involves processing articles already stored in domain specific database (DSDB). These articles currently do not comprise the full text of the original articles, only the sections of documents that are of most interest. These are tables and figure descriptions, referred to as *legends*, which are considered the most probable placeholders for research findings. The full text of the article is not provided as input to the system. It has been established by Telemakus project team that the most interesting information about research literature is usually found in legends [15].



Legend:
Effects of caloric restriction and ad-libitum on young and old age rats .
Syntactic Analysis:
Effects: verb
Of: Preposition
Caloric: noun
Restriction: noun
And: conjunction
Ad-libitum: noun
On: Preposition
Young: noun
And: conjunction
Old: noun
Age: noun
Rats: noun
Trigger word Finding :
If (special trigger sentence){
Trigger = special trigger sentence
}
Else if (preposition){
Trigger = preposition
}
SPECIAL TRIGGER WORD IS PRESENT:
EFFECTS OF
Triple formation :
If(special trigger){
Find the seperator for that trigger word and divide the sentence at the seperator}
If(preposition){
Separate the sentence at the preposition}
SEPERATOR FOR SPECIAL TRIGGER WORD:
ON
TRIPLE: (effects of caloric restriction and ad-libitum , effects of, young and old age rats)
Semantic Analysis:
Phrase 1:
Effects of caloric restriction and ad-libitum
Caloric restriction:
Ad-libitum
Are the concepts in the phrase .
Phrase 2:
Young and old age rats
Young:
Old:
Age:
Rats:
Are the concepts in phrase 2.
Relationship selector:
Relationships are:
(caloric restriction, young)
(caloric restriction, old)
(caloric restriction, age)
(ad- libitum, young)
(ad- libitum, old)
(ad- libitum, age)
Ranker:
caloric restriction, young, 100)
(caloric restriction, old, 100)
(caloric restriction, age, 100)
(ad- libitum, young, 100)
(ad- libitum, old, 100)
(ad- libitum, age, 100)
Evaluation:
Precision:
Recall:

Fig. 3. Concept Miner system process flow

Given an article or a set of articles, the system starts by extracting all legends already stored in the database, processes each legend by identifying interesting relationships, filters relationships, ranks those relationships based on a number of parameters, and finally writes the resulting relationships to an XML file for later use. For comparison purposes, precision and recall are also computed by the system on a per-article basis.

The concept mining process can be divided into three main phases, *syntactic analysis*, *semantic analysis* and *concept mapping and association*.

4.1 Syntactic Analysis

In a given document, there may be one or more legends. Each legend may contain one or more sentences. For any given legend, the text is first broken into constituent sentences. The process of concept extraction and association is applied at the sentence level. Each sentence is parsed and grammatical structures are extracted. From the concept association perspective, each sentence is made up of a connector phrase, called a *trigger phrase*, and the two phrases connected by that trigger phrase. An example of trigger phrase shown on Fig. 3 is "effects of". These trigger phrases are usually prepositions, but human experts have also provided special phrases that act as triggers, such as "effect of". A trigger phrase may contain a connector phrase that separates the remaining part of the sentence into two phrases. After a trigger is found in a sentence, the remaining sentence is split into two phrases optionally connected by a connector phrase. This phase of the system is called *syntactic analysis* in a broad sense. The connector word and two phrases together are called a *triple*. Syntactic parsing determines the structure of the sentence being analyzed. Syntactic analysis involves parsing the sentence to extract information contained within word ordering. Syntactic parsing is computationally less expensive than semantic processing. Syntactic analyzers can be classified into two types depending on their approach to analysis. *Top-down analyzer* starts from the root symbol of the grammar and successively predicts (usually in a left-to-right manner) what its constituent parts should be. On the other hand, *bottom-up analyzer* starts from the string to be analyzed and attempts to construct a syntactic tree by recognizing the right-hand sides of the grammar rules and thus reducing those portions of the string to their non-terminals. This process is repeated until the only remaining non-terminal is the root symbol of the grammar.

For the purpose of this project, the *bottom-up analyzer* is the best fit based on the input pattern, which means that the input string should be taken as input and parsed. Currently, the system uses a basic parser API called *Specialist Text Tools* API that is an open source java implementation [11]. This parser is a minimal commitment barrier category parser. The minimal commitment analysis assigns underspecified syntactic analysis to lexically analyzed input. The parser package contains a shallow parser that extracts minimal phrases from sentences. Using the *Specialist lexicon*, the part of speech and other syntactic information are analyzed. This analysis is specific to biomedical field. The *Specialist Text Tools* tokenizer package tokenizes text into words, sentences, and sections. It can handle free text and Medline citation formats. Sentences are found by looking for sentence bounding punctuation for the most part, and looking at the capitalization of the next word that follows. This method is not always successful, particularly when abbreviations such as *Dr.* and *Mr.* are met. A list of acronyms is consulted when periods are hit. Two new lines in a row are also considered a sentence break. By the end of processing, an analyzed sentence contains all the tokens that make up the sentence, along with their character offsets back to the original document. The results of this phase are a set of *triples* (see Fig. 3 for an example).

4.2 Semantic Analysis

After triples are built, each triple is further analyzed by *semantic analysis*. This involves looking for concepts in each phrase, and is accomplished by applying a domain specific natural language processing tool. From each phrase, a candidate list of concept phrases from the UMLS is extracted. The semantic analysis is made possible by the National Library of Medicine (NLM)'s UMLS [12]. UMLS ultimate goal is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health.

Although the *Specialist Text Tools* also resort to UMLS, it is a foremost knowledge source for semantic analysis. The words or phrases are considered as concepts in the medical domain if said words or phrases can be found in the metathesaurus of the UMLS. Metathesaurus is one of the three UMLS knowledge sources. It is the central vocabulary component of the UMLS [12].

In this project, initially, UMLS metathesaurus entries have been extracted manually in order to create a consistent controlled base vocabulary specific to the domain of caloric restriction and nutritional aspects of aging, and placed within DBSB database. These UMLS selected thesauri characterize mainly research concepts and process description, such as organism type. As new concepts are identified from the documents' tables and figures, they are translated into their UMLS preferred terms, which are added to the controlled vocabulary database.

Semantic analysis is performed on the results of syntactic analysis of the legends to determine the meaning of the words in the sentence. In this step, the semantics of each word or phrase is evaluated. Though there are a few choices for performing semantic analysis on free text, this project uses *MMTx* tool [10] as it is specifically developed for the biomedical field. The main purpose of *MMTx* semantic analysis is to find out the phrases and their variants and then match these to the phrases or words in the UMLS database. The words or phrases successfully mapped to the UMLS database can be considered as concepts in the biomedical or health field. The concept mapping process can be summarized as follows (see Fig. 3 for an example):

- Parse the text into noun phrases and perform the remaining steps for each phrase;

- Generate the variants for the noun phrase where a variant essentially consists of one or more noun phrase words together with all of its spelling variants, abbreviations, acronyms, synonyms, inflectional and derivational variants, and meaningful combinations of these;
- Form the candidate set of all Meta strings containing one of the variants;
- For each candidate, compute the mapping from the noun phrase and calculate the strength of the mapping using an evaluation function. Order the candidates by mapping strength;
- Combine candidates involved with disjoint parts of the noun phrase, recompute the match strength based on the combined candidates;
- Select those having the highest score to form a set of best Meta mappings for the original noun phrase.

4.3 Concept Mapping and Association

Semantic analysis produces a candidate list of concepts for each phrase. This list is refined in multiple steps. First, duplicates or substrings that may be referring to the same concept are removed at the sentence level. For example, two candidate concepts “muscle mass increase” and “muscles” are considered as same candidate concepts. The latter is removed from the list since the algorithm favors the most precise concepts. Next, a known synonym transformation is performed. This step is necessary to replace generic candidate concept names with those preferred by the domain. For example, “free access to food” is replaced with “ad libitum”.

To improve precision, this list is further filtered by pattern matching the concept variations. A preliminary list of relationships is created from these two candidate concept lists of left and right phrases. This is the *relationship selector* phase. General approach to association is based on trigger phrase in a sentence. Each sentence is conceptually equivalent to a list of left-side candidates and a list of right-side candidates connected by a trigger phrase. From these lists, a set of relationships is constructed. Each relationship is a candidate association, containing exactly one left-side concept and one right-side concept (see Fig. 3 for an example).

At the article level, all unique relationships from constituent sentences are aggregated. This list is again refined to remove partial matches. This step is necessary to remove partial matches that may be unique at the sentence level but partially equivalent at the article level.

This list of relationships is ranked based on the importance of concepts, in particular based on the presence of the concepts in the domain-specific database. They are further ranked based on the weight of the relationships involved. The weight of a relationship is the sum of the weights of concepts in that relationship. The weight of a concept is calculated as the number of times this concept occurs in a relationship.

After obtaining a short list of relationships, the effectiveness of the system is evaluated by precision and recall. Finally, the results are written to an XML file, and made available to other parts of the Telemakus system, like Crosscheck.

6 Results

This system is designed to analyze multiple documents at the same time. The success of the system is determined by the recall and precision ratios. Current results show an average recall of 81% and precision of 50% for partial match. Precision is the ratio of matching relations to the total number of relations identified. Recall is the ratio of matching relations to the total number of relations identified by the manual process. The precision and recall are calculated in two ways: partial matching and total matching. In partial matching strategy, if the system extracted relationship (muscle mass – caloric restriction) and the manual results has relationship (muscle mass increase – caloric restriction), then this relationship is considered a match. In total matching, the relationship should be present in the manual results exactly matching both concepts.

Table 1. Precision and recall ratios

Number of Documents	Total Recall	Total Precision	Partial Recall	Partial Precision
Total	**53%**	**35%**	**81%**	**50%**

The system is evaluated for 30 random articles. The average values of recall and precision for these 30 documents are shown in Table 1. It shows that the average values of precision and recall are much higher when partial matches of the concepts are also considered as a match. The reason for considering partial matching is that, there can be some implied knowledge that is used by the domain expert during the manual process, but that kind of knowledge is either not available to this system or hard to automate. These results are encouraging because relationships mining is a much more complex task, in particular when it involves semantic analysis such as in this system, than classical information retrieval. Some of the relationships retrieved do not even share a word with sentences in the documents. These figures compare system performance with human performance, while even humans between themselves would not retrieve the same relationships. An interesting result is that the relationships extracted make good sense, even though the human indexer may not have selected these as the most important in a document.

7 Conclusion

Mining concepts to index literature is a complex task that requires many levels of refinements at both the syntactic and semantic level. Although the results so far are encouraging, several ways are still open for refinements. In the future, the algorithm to extract relationships can be improved in particular by eliminating the legends that represent facts, instead of research findings. Another improvement will be to store the legends in the database in full, so that the input to the automated system and the manual system are the same. This system approach is very original in comparison with other systems mining biomedical literature for concepts, because it mines for relationships between concepts, and not for isolated concepts. Telemakus indexes and navigates through the documents database by relationships between concepts.

Acknowledgements

We want to thank Sherrilynne Fuller, Debra Revere, and Paul Bugni, from the Division of Biomedical and Health Informatics of the University of Washington for providing the idea, the data, and their support role throughout this work.

References

Chang, C., Hsu C.**:** Enabling concept-based relevance feedback for information retrieval on the WWW. Knowledge and Data Engineering (IEEE) vol.11 issue 4 (1999) 595-609

1. Dorre, J., Gerstl, P., Seiffert, R.: Text mining: finding nuggets in mountains of textual data. In: Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM press (1999) 398-401
2. Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A.: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, Bioinformatics 17 Suppl 1 (2001) S74-S82
3. Fuller, S., Revere, D., Bugni, P., Martin, G.M.: A knowledgebase system to enhance scientific discovery: Telemakus. Biomed Digit Libr. Sep 21;1(1):2 (2004)
4. Hand, D., Mannila, H., Smyth, P.: Principles of data mining. MIT Press (2001)
5. Hearst, M.A.: Untangling Text Data Mining. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland (1999) 3-10
6. Jiawei, H., Micheline, K.: Data mining concepts and techniques. 1st edn. Morgan Kaufmann (2000)
7. Lin, S., Chen, M.C., Ho, J., Huang, Y.: ACIRD: Intelligent Internet Document Organization and Retrieval. IEEE Transactions on Knowledge and Data Engineering vol. 14 (2002) 599-614
8. Nasukawa T., Nagano, T.: Text Analysis and Knowledge Mining System. Knowledge management Special Issue. IBM systems journal Vol. 40 (2001) 967-984
9. National Library of Medicine: MetaMap Transfer (MMTx). (2005) http://mmtx.nlm.nih.gov [Last access: 2005-04-01]
10. National Library of Medicine: The Specialist NLP Tools. (2004) http://specialist.nlm.nih.gov [Last access: 2005-04-01]
11. National Library of Medicine: The Unified Medical Language System. (2005) http://umls.nlm.nih.gov [Last access: 2005-04-01]
12. Swanson, D.R.: Information discovery from complementary literatures: Categorizing viruses as potential weapons. Journal of the American Society for Information Science Vol. 52(10) (2001) 797-812
13. Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence Vol.9 (1997) 183-203
14. Telemakus project team: Mining and Mapping Research Findings to Promote Knowledge Discovery (2001) http://www.telemakus.net/papers.html [Last access: 2005-04-01]