Model of a heuristic knowledge organizing system for searching and classifying structured data

Hristina Moneva
Assistant Professor
Department of Computer Systems and Technologies,
University of Veliko Turnovo "St. Cyril and St. Methodius"
2 Teodosii Tarnovsky Str.
5003 Veliko Tarnovo, Bulgaria
+359 62 649831
xmoneva@ieee.org

Margarita Todorova
Associate Professor, Head of Department
Department of Computer Systems and Technologies,
University of Veliko Turnovo "St. Cyril and St. Methodius"
2 Teodosii Tarnovsky Str.
5003 Veliko Tarnovo, Bulgaria
+359 62 649831
marga_get@abv.bg

ABSTRACT
Different approaches for searching, classifying and visualization data are presented. The generalized model of a heuristic knowledge organizing system for searching and classifying a structured data is presented, as well as its basic components. The user's roles with their characteristics, interests and activities, classifying, searching and result sub systems in the process of functioning are pointed.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval –*Information filtering; Relevance feedback; Retrieval models; Search process.*

General Terms
Management, Measurement, Design.

Keywords
searching, classifying, fuzzy sets, SOM.

1. INTRODUCTION
Advantages and disadvantages of classification [3]:

A site that organizes knowledge with a classification scheme demonstrates several advantages over sites which do not:

- Browsing: classified subject lists are easily able to be browsed in an online environment. Browsing is particularly helpful for inexperienced users or for users not familiar with a subject and its structure and terminology;
- Broadening and narrowing searches: classification schemes are hierarchical and therefore can be used to broaden (i.e. for improved recall) or narrow a search when required;
- Context: the use of a classification scheme gives context to the search terms used;
- Potential to permit multilingual access to a collection: since classification systems often use notations independent from a specific language, indices in different languages can offer multilingual access to the same resources without any further changes to the collection;
- The partitioning and manipulation of a database: large classified lists can be divided logically into smaller parts if required;
- The use of an agreed classification scheme could enable improved browsing and subject searching across databases;
- An established classification system is not usually in danger of obsolescence. Some classifications may have to be changed when a new edition of a scheme is published, but it is unlikely that every single resource will have to be re-classified;
- They have the potential to be well-known: regular users of libraries will be familiar with at least part of one or more of the traditional library schemes;
- Many classification schemes are available in machine-readable form.

Classification schemes, however, can be sometimes subject to criticism:

- The division of logical collections of material: classification schemes often split up collections of related material;
- The illogical subdivision of classes: some popular schemes do not always subdivide classes in a logical manner;

Assimilating new areas of interest: classification schemes, since they are usually updated through formal processes by organized bodies, often reveal difficulty in reacting to new areas of study.

2. CLASSIFICATION SCHEMES

2.1 Types of classification process:
- Manual (human-powered);
- Automatic (machine-powered).

2.2 Classification schemes can be defined by several categories, but can be broadly divided into [3]:
- Universal schemes – standardized schemes;
- National general schemes - universal in subject coverage but usually designed for use in a single country;
- Subject specific schemes - designed for use by a particular subject community;

• Home-grown schemes - schemes devised for use in a particular service.

The type of classification scheme chosen for use in an Internet service should depend upon the scope of service which is planned.

2.3 Current use of classification schemes in existing search services [3]:

• Extent of usage in Internet services;

• Extent of usage in traditional and other online services;

• Multilingual capability;

• Strengths and weaknesses of the scheme;

• Integration between classification scheme and other systems e.g. controlled subject headings;

• Linking DDC to third party classification data;

• Digital availability;

• Copyright;

• Extensibility and development effort of the scheme;

• Other issues.

3. SEARCH METHODS

3.1 General paradigms

3.1.1 Brute-force/Blind search

Also called weak search methods, most general methods are brute-force because they do not need domain knowledge; however, they are less efficient as a result. [5]

3.1.2 Heuristic search

Heuristic searches use some function that estimates the cost from the current state to the goal, presuming that such a function is efficient. Generally speaking, heuristic search incorporates domain knowledge to improve efficiency over blind search. [5]

3.2 Request types [13]

• Boolean search: A search allowing the inclusion or exclusion of documents containing certain words through the use of operators such as AND, NOT and OR;

• Concept search: A search for documents related conceptually to a word, rather than specifically containing the word itself;

• Full-text index: An index containing every word of every document cataloged, including stop words (defined below);

• Index: The searchable catalog of documents created by search engine software. Also called "catalog". Index is often used as a synonym for search engine. Index is commonly pluralized as "indices." However, Search Engine Watch instead uses the alternative plural form "indexes.";

• Keyword search: A search for documents containing one or more words that are specified by a user;

• Phrase search: A search for documents containing an exact sentence or phrase specified by a user;

• Precision: The degree in which a search engine lists documents matching a query. The more matching documents that are listed, the higher the precision. For example, if a search engine lists 80 documents found to match a query but only 20 of them contain the search words, then the precision would be 25%;

• Proximity search: A search where users to specify that documents returned should have the words near each other;

• Query-By-Example: A search where a user instructs an engine to find more documents that are similar to a particular document. Also called "find similar";

• Recall: Related to precision, this is the degree in which a search engine returns all the matching documents in a collection. There may be 100 matching documents, but a search engine may only find 80 of them. It would then list these 80 and have a recall of 80%;

• Relevancy: How well a document provides the information a user is looking for, as measured by the user.

4. RESULT VISUALIZATION

4.1 Catalog

The paradigm of browsing a directory of topics arranged in a tree where children represent specializations of the parent topic is now pervasive. The average computer user is familiar with hierarchies of directories and files, and this familiarity carries over rather naturally to topic taxonomies.

Topic directories offer value in two ways. The obvious contribution is cataloging of Web content, which makes it easier to search. The second contribution is in the form of quality control. [4]

4.2 Sorted list of documents

Short queries, unless they include highly selective keywords, tend to be broad because they do not embed enough information to pinpoint responses. Such broad queries matched thousands to millions of pages, but sometimes missed the best responses because there was no direct keyword match. [4]

4.3 Visual display

Visual methods are used to display data in ways that capitalize upon the particular strengths of human pattern processing abilities. [6]

5. GENERALIZED MODEL OF A HEURISTIC KNOWLEDGE ORGANIZING SYSTEM [12]

5.1 Components of the presented model

There are three basic components in the presented model. Their relations and behaviors specify system functionality. These components are process participators, classifying and searching sub systems.

5.1.1 Process participators

In the working process of the system the following user types are outlined:

• Expert – enters his knowledge as part of the system;

- User – uses the system resources aiming to receive the needed knowledge.

According to the system's normal functioning it is needed:

- Administrator – assists the system's functioning and gives the corresponding access rights to the different user types.

5.1.2 Classifying subsystem

The classifying sub system has two basic processes which can characterize it – data presentation and classifying. For their development different approaches can be used and thus can lead to different functionality and result.

5.1.2.1 Data presentation:

In real world the relations between objects not always can be presented in conventional mathematical way – with pure false or true, 0 or 1. It is needed their real relation to be presented. For that reason fuzzy set theory and its apparatus can be used.

The fuzzy set is defined with some base scale B and membership function μ(x), $x \in B$ which values are into [0,1]. Consequently the fuzzy set B is a collection of ordered pairs (x, μ(x)), where $x \in B$. [2]

When the membership function gives the subjective expert level, the concrete value of the base scale is the definition of a fuzzy set. This function is not a probabilistic that has an objective character and does not follow any other mathematical dependencies.

In this way it can be presented more clear and adequate how the set of the documents arranged in hierarchical structure of domains or sub domains are characterized from the keywords set.

The fuzzy sets give possibilities to define the subjective opinion of different individuals.

5.1.2.2 Classifying:

In the context of data classifying, fuzzy sets can give more clear and adequate way for document presenting with their keywords which characterize it and its membership function (level of adequacy [11]), which define their relevancy to the document. In publishing the document, the keywords are manually set up by the author. It is assumed that he is more involved in the matter of the document and can define keywords and their level of adequacy more correct than anyone else. [10]

Let assume that the classifying structure will be the universe set **U** containing all possible thematic domains – all domains and their sub domains $\mathbf{A_1,A_2...A_n}$. The system is aiming at being comprehensive by accumulating more and more documents, but in real life nothing is absolute and thorough. For that reason every expert during publishing in the system is allowed to add new domains and sub domains if it is needed.

Let the elements of set **D** are the documents $\mathbf{d_1,d_2...d_m}$.

Let the set **R** be the binary relation from a set **U** to a set **D**, which gives information about classification of documents into domains and sub domains. The relation **R** is described by characteristic function of set **R** - $\left\|\mu_{Rij}(A,d)\right\|$, i=1,2...n, j=1,2...m,

define as $\mu_{Rij}(A,d) = \begin{cases} 1, if\ (A_i, d_j) \in R \\ 0, if\ (A_i, d_j) \notin R \end{cases}$ [1].

Let **K** be the set with elements $\mathbf{k_1,k_2...k_l}$ – keywords describing some document of the set **D** with the fuzzy relation **S** defined with $\mu_S(d,k) \in [0,1]$.

5.1.3 Searching subsystem

The searching sub system has two basic processes which can characterize it – the request that the user send to the system and the different types of result system answers as response and which may lead to concrete result list of documents or may give new possibilities for searching.

5.1.3.1 Request:

When the user is searching he sends to the system request as combination of keyword and level of adequacy. The searched *level of adequacy* is a number into [0,1] which is set by the user and characterizes the relevancy between searched keyword (phrase) and proposed documents by the system.

Let the searched word be **ks**. Let the searched level of adequacy to this word be $\boldsymbol{\alpha_{ks}}$. Consequently the given level of adequacy is the level $\boldsymbol{\alpha_{ks}}$ of the fuzzy relation **S** of **DxK**, respectively $\boldsymbol{\mu_S(d,k) \geq \alpha_{ks}}$.

A complex request is a request that has words such as "and", "or", "not" and are corresponding to fuzzy sets operations – "conjunction", "disjunction" and "complement".

They can be summarized in:

AND	$\wedge$ (conjunction)	min
OR	$\vee$ (disjunction)	max
NOT	$\bar{k}_i$ (complement)	1 - k_i

By representing the searched word in this way we can find the fuzzy set **DS** which contains "all documents with the keyword **ks**" and where the level of adequacy is set to $\boldsymbol{\alpha_{ks}}$.

5.1.3.2 Searching:

A few types of searching can be differentiated:

- By keywords and level of adequacy – The level of adequacy measures the relevancy of the searched word and retrieved documents as result. With higher level of adequacy the number of result documents is reduced. This way in advance is dropping away some documents not enough relevant to the concrete problem. The structure of domains and sub domains where relevant documents are found is presented as a result. Thus for the second time some documents are dropped away because the user looks only in the domain he is interested in.

- By statistics of a concrete domain or sub domain – there have two different types: by evaluation measures and by list of the most common used and with higher level of adequacy to the domain.

- Visual – map of some domain which visual gives the notion of the most specific and the most common keywords. This is alternative of searching by statistic of a domain. By this way the user can obtain new kind of information – he can see more keywords relevant to the searched one or to find the most specific keyword and their relation aiming at improving his request. This is especially suitable for users that are not so familiar with the essence of the domain.

5.1.3.3 Presenting the result in domains:

- By evaluation measures:

The general case is possible as one keyword describes documents of more than one domain and their sub domains, and / or as one document has been classified in more than one domain or sub domain. Consequently when is searched for such a keyword, the result will be two or more different hierarchies grouping the result thematically in domains and their sub domains.

The results retrieved as a response of request, may be presented as a hierarchical structure of domains characterized by parameters result depending.

On one side may be used the interval where results belong to:

[level_of_adequacy÷optimistic_expectation] or

$A_i[\alpha_{ks} \div \max \min \mu_{R \circ S}(A_i, ks)]$.

On the other side – two characteristics of the evaluation techniques used in IR, presented as values in percents or using their numerical components. These characteristics are:

$$\text{precision} = \frac{(\text{number of relevant items retrieved})}{(\text{total number of items retrieved})},$$

$$\text{recall} = \frac{(\text{number of relevant items retrieved})}{(\text{number of relevant items in collection})}$$ [8].

The result may be presented like:

Domain [level_of_adequacy÷optimistic_expectation / precision_% / recall_%]

or

Domain [level_of_adequacy÷optimistic_expectation / number_of_relevant_items_retrieved / total_number_of_items_retrieved / number_of_relevant_items_in_collection]

- By statistics of a concrete domain or sub domain – a list of the most common used and with higher level of adequacy to the domain is presented like:

Domain:

keyword_1 [min÷average÷max, number_of_relevant_items_retrieved / total_number_of_items_retrieved],

keyword_2 [min÷average÷max, number_of_relevant_items_retrieved / total_number_of_items_retrieved],

...

keyword_n [min÷average÷max, number_of_relevant_items_retrieved / total_number_of_items_retrieved].

5.1.3.4 Presenting the result list of documents:

After choosing a concrete domain to user is presented an ordered list of documents which is descending by their relevancy to the request. They can be visualized like:

Domain_ M /... / Sub domain_ N:
Document_No1 (level_of_adequacy)
keyword_1 (#);kw_# (#);kw_# (#);…
Document_No2 (level_of_adequacy)
keyword_1 (#);kw_# (#);kw_# (#);…

5.1.3.5 Visualization by SOM-based method:

There have a few reasons that led to the development of SOM-based methods for data exploration – limitations of data mining by search engines, multiple sources and formats of information, and measuring similarities often suffices. [9]

"The SOM is an unsupervised-learning neural-network method that produces a similarity graph of input data. It consist of a finite set of models that approximate the open set of input data, and the models are associated with nodes (neurons) that are arranged as a regular, usually 2-D grid. The models are produced by a learning process that automatically orders them on the 2-D grid along with their mutual similarity." [7]

The SOM algorithm can be used to generate a graphical display of the document collection after all the documents have been encoded as numerical vectors. The construction of a large SOM of a large document collection is a computationally intensive task [9]. To avoid this problem, this algorithm can be used for preliminary reduced number of documents. This will make the task less intensive and the result will become faster.

Due to the SOM method there are some basic stages according to document encoding [9]. Preprocessing – removing the parts that are not considered relevant for the organization. This is because of the automatic extraction of keywords. Vector space model – the frequency of occurrence of each word in a document is computed and all frequencies are collected into a vector that produces a vector space model. In the proposed model the weighs are based on the level of adequacy that is given by the expert. Semantic relations – the vector space model does not take into account semantic relations between the words but SOM – does. Dimensionality reduction – the vector space model is effective for small document collections but not for large ones. To solve this problem random mapping method is used which produce almost as good results as the original vector space model. It consists of a multiplication of the original document vector with a random matrix that produces a smaller output dimensionality. Incorporating auxiliary knowledge – may be used thesauri to find semantic relations between words or linguistic algorithms for finding the stems of the words.

6. CONCLUSIONS

The proposed model gives the possibility to classify documents adequately in advance according to their authors which are experts in the specific domain not only with keywords but with their specific level of adequacy when classifying; and not only giving a result list of documents but to use self organizing map, to give more information about the relations between the keywords in concrete area when searching.

7. REFERENCES

[1] Бърнев П., П. Станчев (1987). Размити множества. София: Народна просвета,. 110с.

[2] Гаврилова Т. А., В. Ф. Хорошевский, (2000), Базы знаний интеллектуальных систем, Санкт-Петербург: Питер.

[3] Brümmer, A., Hiom, D., Peereboom, M., Poulter, A., Worsfold, E.,. The role of classification schemes in Internet

resource description and discovery. Work Package 3 of Telematics for Research project DESIRE (RE 1004), http://www.ukoln.ac.uk/metadata/desire/classification/class_tc.htm

[4] Chakrabarti, S. Mining the Web – Discovering knowledge from hypertext data. Elsevier Science, 2003.

[5] Doyle, P. Search Methods. http://www.cs.dartmouth.edu/~brd/Teaching/AI/Lectures/Summaries/search.html

[6] Hand, D., Mannila, H., Smyth, P. Principles of Data Mining. MIT Press, 2001.

[7] Kohonen, T., S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela (2000). Self organization of a massive text document collection. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, vol. 11, pp. 574-585.

[8] Lagus, K. (2000). Text retrieval using self-organized document maps, Technical Report A61, Helsinki University of Technology, Laboratory of Computer and Information Science.

[9] Lagus, K., T. Honkela, S. Kaski, T. Kohonen (1998). WEBSOM for textual data mining. Netherlands: Kluwer Academic Publisher.

[10] Lamont, J. (2005). Unlocking enterprise data: Metadata holds the key. KMWorld – April 2005.

[11] Moneva, H., (2004), Model of an intellectual search engine represented by the fuzzy sets theory, Rousse: CompSysTech2004.

[12] Moneva, H., Todorova, M., Model of hybrid system for searching and classifying a structured data. ECET, Berlin, 2005.

[13] Sullivan, D. Search Engine Glossary. http://searchenginewatch.com/facts/article.php/2156001