IMPROVEMENT OF SEARCH EXPERIENCE BASED ON MESH SEMANTICS: RECENT INNOVATIONS IN MEDVIK PORTAL

Filip Kriz, Ondrej Horsak, Lenka Maixnerova, Helena Bouzkova, Eva Lesenkova, Adela Jarolimkova

National Medical Library, Prague, Czech Republic
(filip.kriz@gmail.com)

Introduction

National Medical Library (NML) of the Czech Republic operates Medvik Portal - a web application for access to several bibliographic databases produced by NML and cooperating institutions. The main databases are Medvik Catalogue - describing library collections with the holdings of the Union Catalogue of Czech Medical Libraries, and Bibliographia medica Czechoslovaka (Bibliomedica) - Czech national medical bibliography. The total amount of bibliographic records approximates 700 thousands.

The technical design of Medvik Portal with multiple underlying databases was not sufficient for fast and reliable retrieval, which has led to further development. An aggregated database specially optimized for searching tasks has been created. The database is updated automatically from production databases and allows efficient access to all bibliographic data from one access point using full-text search approach. This design allows a Google-like search experience, but there is a need to add more functionality to achieve better precision and recall. This need comes from obvious limitations of rather unfocused full-text searching and also from our users' feedback and recent NML's survey.

NML translates Medical Subject Headings (MeSH) into Czech language using MeSH Translation and Maintenance System (MTMS). Majority of our bibliographic records are indexed with MeSH descriptors, qualifiers and NML Subject terms which all can be used for concept-based and context-sensitive tools for a better search experience and more pertinent results.

We have focused on the MeSH semantics to develop following features: search term suggestion and query builder, multiple clustering of search results, tag clouds for query representation and browsing support, and the Subject browser - aggregating MeSH, NML Subjects and Supplementary Concepts Records. The techniques used to implement these features will be described further.

Background

The design of Medvik Portal (http://www.medvik.cz) has not changed much since its launch in 2006. Though several features have been added during past years based on our users' feedback. NML has also implemented new systems and services that needed to be integrated. The Medvik system is further described in [1]. It had been clear that a major redesign or upgrade was necessary to catch up with new developments in modern search interfaces design.

The fact that medical professionals have not been very satisfied with the current interface also showed in the results of the last year NML's survey called "Information behaviour of medical professionals in relation to modern library and information services in health care". In the evaluation section of the portal, the interface and navigation was rated 2.2, and graphical design 2.6 on 1 to 5 point scale (good-bad) as can be seen in figure 1.

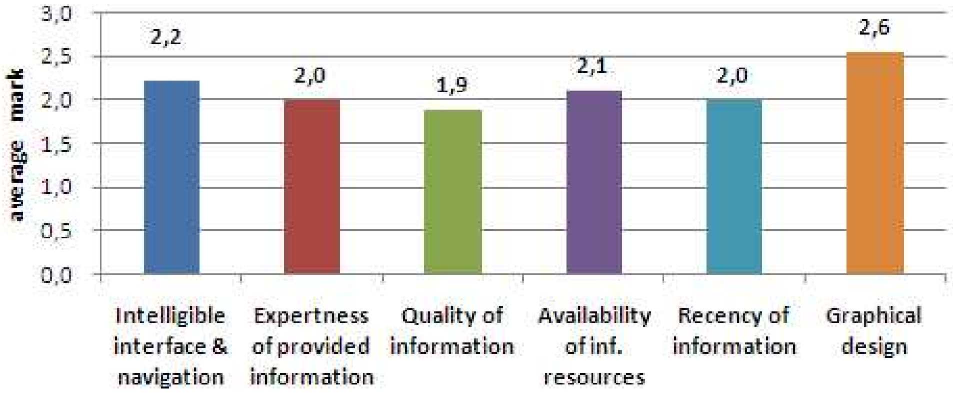


Fig. 1. Medvik Portal evaluation results

These needs have led to efforts to make Medvik Portal more user-friendly and accessible so the portal would provide a better user experience. Initial effort was focused on the integration of our production databases into one “index” database to allow simultaneous and efficient searching of all NML’s datasets. The database contains mainly data elements that are suitable for retrieval purposes and its structure is optimized for full-text searching. The database content is additionally populated with more selecting terms from MeSH and is automatically updated on regular basis. The data structures are independent of the MARC format used in production databases and thus are more flexible for the interface development.

We have come to a conlusion that our scenario is similar to PubMed: the majority of bibliographic records are indexed using MeSH headings and the Bibliomedica article database is in scope comparable to Medline. There are many PubMed alternative interfaces and results post-processing tools available on-line [2,3] which use interesting features: query building aids, visualization and refinement of results, related documents displays etc. These features allow users to choose selection terms for building a query, to select sets of documents of interest from visual results display, to discover other related documents and information relevant to their needs. We have analyzed the methods used, in particular by LigerCat [3] and Anne O’Tate [4], and we have tried to implement them with some modifications explained further.

Methods and implementation

The value of using MeSH descriptors in retrieval process has been proved [5]. On the other hand, most users are not familiar, due to different reasons, with MeSH thesaurus and its usage in retrieval. Thus we have designed the basic searching workflow as follows: users can start with a simple text search which is performed against our index database. Because of the full-text enabled search, they are likely to receive more (or at least some) results that can be further explored or expanded using several clusters, tag cloud and filters. By simply clicking on the MeSH headings and other cluster terms, they can browse and refine the results and the selected terms appear in the query builder.

Users can at any time re-run the query with the terms selected or apply filters and thus receive more accurate results. Users have the option to remove or exclude a term from the query before the re-run. The selected descriptors used in the query are automatically exploded that records with narrower (more specific) terms are retrieved as well. The scheme of the workflow is presented in figure 2. Users can select document records from the results display for later actions or take a look at a detailed document view.

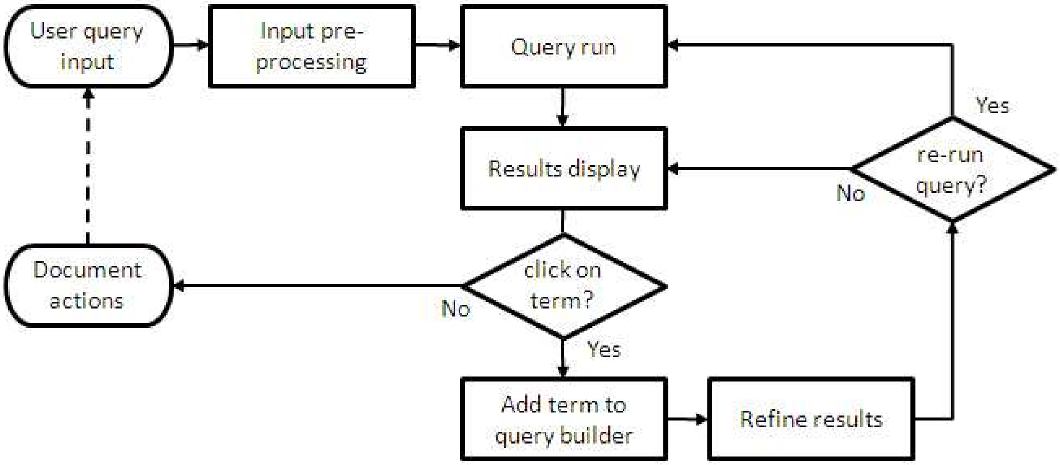


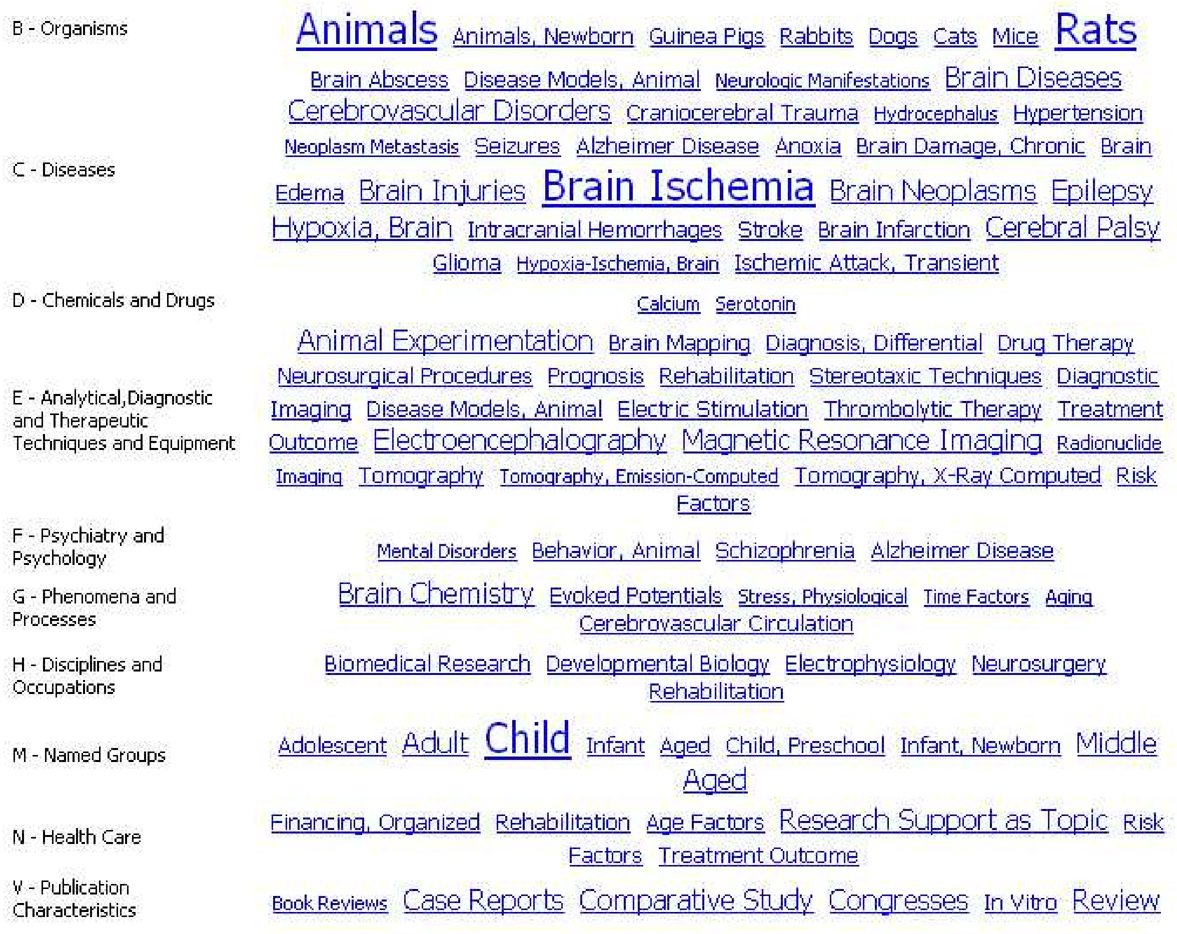
Fig. 2. Scheme of the basic search workflow

Clustering analysis (clustering) is a common approach to show the similarity between a set of observations. Clustering of PubMed/Medline records is usually performed by vector space model [6] based on documents content similarity (words from titles and abstracts). Another approaches are based on MeSH semantics (that two documents indexed with the same or related MeSH descriptors are similar) or combining both content and semantic similarity. We decided to work with the descriptors and MeSH data structures only to enable the clustering and tag cloud features in first version of the new interface which would make the initial development less complex. Also the availability of abstracts in Bibliomedica records is low - there are 12 000 Czech and 10 000 English abstracts from the total of 570 000 bibliographic records.

The MeSH data (MeSH descriptors and concepts, qualifiers and Supplementary Concept records) have been imported into SQL database from XML files provided by National Medical Library, Bethesda, using SQL XML Bulk Load. We have performed minor changes in MeSH data structures that Czech and English equivalent terms appeared on same table rows. This design allows to easily switch between the two language versions in the interface and to search MeSH terms independently of the language used. The tree structure has been implemented using the nested sets technique.

Tag clouds emerged on the Web in 2006, first in photo sharing service Flickr [7] and they shortly became common feature of many Web 2.0 applications (Del.ici.ous, Technorati, Connotea, blogs etc.) The tags in these applications are usually user-generated in the process of collaborative (or social) tagging resulting in classification known as folksonomy. The tag clouds are a form of visual retrieval interfaces allowing to browse collections of web resources. The tag terms are weighted based on their frequency in a given collection and the more frequent (important) tags are displayed in bigger fonts. The tags in a cloud are often sorted alphabetically though it is not always the best option. There have been attempts to cluster the tags semantically that tags close in meaning appear close to each other [8].

We have adopted the later approach to create a cluster of MeSH descriptors based on their frequency in the resulting records. The calculation of the frequency and final weight of terms is inspired by LigerCat and AnneO'Tate algorithms [3,4]. No descriptors are excluded from the cloud except the MeSH Check tags (Male, Female, Adult etc.) and the Publication types which are accessible in the refine section of the layout. The terms are then mapped to the MeSH Tree structure and grouped into16 top-level categories (the Publication Characteristics branch and check tags are omitted). The terms in each category form a row (or block of tags) and they are sorted by their position in the tree structure. This design enables to show semantically related terms (in the same tree branch) near each other. Thus users can easily navigate the categories and move from general to more specific terms that might be of their interest. There is also the option to switch on the alphabetical sorting inside the blocks. The cloud is positioned at the bottom of the results page. The example of a tag cloud is depicted in figure 3. The descriptors which reside in multiple branches of the tree appear repeatedly in each category.



B - Organisms	Animals Animals, Newborn Guinea Pigs Rabbits Dogs Cats Mice Rats
C - Diseases	Brain Abscess Disease Models, Animal Neurologic Manifestations Brain Diseases Cerebrovascular Disorders Craniocerebral Trauma Hydrocephalus Hypertension Neoplasm Metastasis Seizures Alzheimer Disease Anoxia Brain Damage, Chronic Brain Edema Brain Injuries Brain Ischemia Brain Neoplasms Epilepsy Hypoxia, Brain Intracranial Hemorrhages Stroke Brain Infarction Cerebral Palsy Glioma Hypoxia-Ischemia, Brain Ischemic Attack, Transient
D - Chemicals and Drugs	Calcium Serotonin
E - Analytical,Diagnostic and Therapeutic Techniques and Equipment	Animal Experimentation Brain Mapping Diagnosis, Differential Drug Therapy Neurosurgical Procedures Prognosis Rehabilitation Stereotaxic Techniques Diagnostic Imaging Disease Models, Animal Electric Stimulation Thrombolytic Therapy Treatment Outcome Electroencephalography Magnetic Resonance Imaging Radionuclide Imaging Tomography Tomography, Emission-Computed Tomography, X-Ray Computed Risk Factors
F - Psychiatry and Psychology	Mental Disorders Behavior, Animal Schizophrenia Alzheimer Disease
G - Phenomena and Processes	Brain Chemistry Evoked Potentials Stress, Physiological Time Factors Aging Cerebrovascular Circulation
H - Disciplines and Occupations	Biomedical Research Developmental Biology Electrophysiology Neurosurgery Rehabilitation
M - Named Groups	Adolescent Adult Child Infant Aged Child, Preschool Infant, Newborn Middle Aged
N - Health Care	Financing, Organized Rehabilitation Age Factors Research Support as Topic Risk Factors Treatment Outcome
V - Publication Characteristics	Book Reviews Case Reports Comparative Study Congresses In Vitro Review

Fig 3. Example of tag cloud based on query “brain” (Publication types and Check tags included)

The clusters for navigating the results (faceted navigation) are situated in the left side column of the interface. There are several blocks: MeSH clusters - headings (descriptors), check tags, publication types and qualifiers; NML Subjects cluster; authors cluster, date of publication cluster and the source (journal) cluster. All clusters are created using the term frequency in the result set and sorted by descending order. Users can browse the clusters from more to less frequent and display associated documents records. The descriptor cluster is based on the similar algorithm as the tag cloud, but is aimed at drill-down browsing of clustered documents and allows adding more specific descriptors to the query builder. The top clustering level is formed by the headings from the second level of the MeSH tree (the so called running heads).

The Subject browser has been designed to help users to identify topics of interest and to choose headings for building a query. The browser allows free text searching in all descriptor, NML subject and Supplementary (chemical) concept records. Terms corresponding to a query are displayed in the left column and when clicked, the record is displayed. All headings (descriptors) with allowable subheadings (qualifiers) can be added to the query builder. Users can specify, before running a query, whether to explode the terms, to find all or any terms, and to limit the query to articles only. This query builder provides the same functionality as the builder in the results page. There is also the possibility to send the created query to the official PubMed interface at NCBI website.

Conclusions

The new portal interface is still under development and as of this writing is not yet publicly available. Before the official launch later this year, we plan to use internal as well as external focus groups which will help us to test and evaluate the interface in order to improve it. The new version will also run in parallel with the current version for a period of 2 months. There are more features which we want to encompass (related documents display, search term suggestion etc.), but further testing is needed before its implementation. Useful enhancement would also be more sophisticated pre-processing of user's initial input but there are no program libraries standardly available that are able to reliably process Czech natural language. The pre-processing might also be used for search suggestion and type-in error correction tool. We plan to implement it in future through collaboration with Czech linguistic professionals. Also, the automated identification of the important words from titles and article abstracts would allow to create better clustering and might help to overcome some imperfections in the MeSH manual indexing. In future, we would like to create data mining and visualization tools for biblometric analysis of Bibliomedica data and to be able to present broader records contexts for answering "what research, who with whom, where, when" types of questions. We use rather practical then theoretical approach in this work, though we hope the new interface would be available soon and it will provide superior service to all interested users.

References

1. Kriz F, Horsak O, Maixnerova L, Bouzkova H. Medical Virtual Library (MEDVIK) - collaborative environment for innovative information services in Czech Republic. In: Towards a New Information Space - Innovations and Renovations : programme, abstracts. 11th European Conference of Medical and Health Libraries; 2008 Jun 23-28; Helsinki, Finland. Helsinki: European Association for Health Information and Libraries; 2008. p. 21-22 [cited 2010 Apr 28]. Available from: http://www.terkko.helsinki.fi/bmf/EAHILpapers/Filip_Kriz_paper.pdf

2. Kaenel I, Iriarte P. Alternative interfaces for PubMed searches. EAHIL Workshop Krakow. 2007 [cited 2010 Apr 28]. Available from: http://www.eahil.net/conferences/krakow_2007/www.bm.cm-uj.krakow.pl/eahil/proceedings/poster/kaenel%20iriarte.pdf

3. Smalheiser N, Zhou W, Torvik V. Anne O'Tate: a tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008;3:2 [cited 2010 Apr 28]. Available from: http://www.j-biomed-discovery.com/content/3/1/2

4. Sarkar IN, Schenk R, Miller H, Norton CN. LigerCat: using "MeSH Clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. AMIA Annu Symp Proc. 2009 Nov 14;2009:563-7 [cited 2010 Apr 28]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815376

5. Chang AA, Heskett KM, Davidson TM. Searching the literature using Medical Subject Headings versus text word with PubMed. Laryngoscope. 2006;116:336–240

6. Zhu S, Zeng J, Mamitsuka H. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics. 2009 Aug 1;25(15):1944-51

7. Bausch P, Bumgardner J. Make a Flickr-style tag cloud. In: Flickr hacks. O'Reilly Press; 2006. p. 82-86

8. Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006. Merida, Spain. October 25-28, 2006 [cited 2010 Apr 28]. Available from: http://www.nosolousabilidad.com/hassan/improving_tagclouds.pdf