2014-05-27

Multilingual query expansion in the Svemed+ bibliographic database : a case study

Gavel, Ylva; Andersson, Per-Olov

J Inf Sci. 2014 Jun;40(3):269-80.
http://doi.org/10.1177/0165551514524685
http://hdl.handle.net/10616/42071

JIS Accepted for Publication
By the Journal of Information Science: http://jis.sagepub.co.uk

Article

Journal of Information Science
1–12

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Multilingual query expansion in the SveMed+ bibliographic database: A case study

Ylva Gavel
Karolinska Institutet University Library, Stockholm, Sweden

Per-Olov Andersson
Karolinska Institutet University Library, Stockholm, Sweden

Abstract
SveMed+ is a bibliographic database covering Scandinavian medical journals. It is produced by the University Library of Karolinska Institutet in Sweden. The bibliographic references are indexed with terms from the Medical Subject Headings (MeSH) thesaurus. The MeSH has been translated into several languages including Swedish, making it suitable as the basis for multilingual tools in the medical field. The data structure of SveMed+ closely mimics that of PubMed/MEDLINE. Users of PubMed/MEDLINE and similar databases typically expect retrieval features that are not readily available off-the-shelf. The SveMed+ interface is based on a free text search engine (Solr) and a relational database management system (Microsoft SQL Server) containing the bibliographic database and a multilingual thesaurus database. The thesaurus database contains medical terms in three different languages and information about relationships between the terms. A combined approach involving the Solr free text index, the bibliographic database and the thesaurus database allowed the implementation of functionality such as automatic multilingual query expansion, faceting and hierarchical explode searches. The present paper describes how this was done in practice.

Keywords
bibliographic databases; multilingual retrieval; query expansion; thesauri

1. Introduction

Karolinska Institutet (KI) in Sweden is one of the largest medical universities in Europe. Being one of the first MEDLARS centers to be established outside the USA [1-4], KI has a long-standing experience in the provision and development of databases. The MEDLARS services were operated by the Medical Information Center (MIC), subsequently to be integrated with the University Library of Karolinska Institutet (KIB). The services provided by MIC included access to MEDLINE, the literature database produced by the National Library of Medicine (NLM) in the USA [5]. MIC was contracted to index Scandinavian journals for MEDLINE. MEDLINE is now available via PubMed and various commercial platforms. Although the MEDLARS system operated by MIC has long since been closed down, KIB still conducts activities related to literature databases.

MIC developed Swemed, a literature database covering Scandinavian journals in the medical field. The name of the database has since been changed to SveMed+. The database structure closely mimics that of MEDLINE. MIC also initiated the Swedish translation [3] of the Medical Subject Headings (MeSH), the controlled vocabulary of MEDLINE. Over the years, the database containing the Swedish MeSH translation has been integrated into various search tools developed at KIB and elsewhere. An algorithm for multilingual query expansion against SveMed+ using the Swedish MeSH database as a backend is described below.

Corresponding author:
Ylva Gavel, Karolinska Institutet University Library, Fe 200, 171 77 Stockholm, Sweden
Ylva.Gavel@ki.se

1.1. The SveMed+ database

SveMed+ contains biblio[Accepted for Publication] Some of the records contain links to full tex[By the Journal of Information Science: http://jis.sagepub.co.uk] is English, Swedish, Norwegian or Danish. Due to linguistic differences, journals in the other major Nordic languages (i.e., Icelandic and Finnish) are not covered. English translations are provided for the article titles. The references are indexed with terms from the MeSH thesaurus.

Originally running under ELHILL, the search system developed by the NLM for the provision of MEDLINE [6-8], SveMed+ has been available through several platforms including Ovid, BRS/Search and Solr. The present paper describes the Solr based version[1], including the data input system behind the Solr index.

1.2. The MeSH

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary with a medical scope [9-12]. Originally developed for indexing of MEDLINE [13-14], it is presently in use in many search systems including various library catalogues, CISMeF [15], Euroethics [16] and SveMed+. The MeSH has been translated into several languages [3,17], making it suitable as the basis of multilingual search tools in the medical field [15,18]. For example, the Swedish MeSH is available as a browsing tool[2].

MeSH indexing involves supplementing the reference of a journal article with terms picked from the thesaurus. The indexer manually selects a number of terms that describe what the article is about. Since the terms derive from a controlled vocabulary, articles covering the same subject are indexed with the same term which means that they can be retrieved in a consistent way.

The preferred terms in the MeSH thesaurus go by various names such as main headings, subject headings, or descriptors. There is also an entry term vocabulary consisting of synonyms and quasi synonyms that can be utilized for synonym mapping [12]. A MeSH term may be combined with a subheading (qualifier) that indicates a specific aspect of the subject at hand.

MeSH terms are arranged in a hierarchical tree structure with broad (general) terms at the root level and successively narrower (more specific) terms in the branches. The tree structure is poly-hierarchical, which means that a term can reside in more than one position in the tree structure. Moreover, a term may have different narrower terms depending on the position in the tree, since different positions reflect different contexts [12].

The positions in the tree structure are denoted by tree numbers. A tree number is a string consisting of alphanumeric character sequences separated by punctuation marks. The parts between the punctuation marks denote categories and levels in the tree structure. Each tree number contains the tree number of the broader parent term as a substring.

When picking terms for indexing, the indexer selects the narrowest (most specific) terms possible. A user submitting a query for a subject corresponding to a certain MeSH term will usually also want to retrieve articles indexed with narrower terms. Otherwise, a loss in recall will occur. Entering all the terms explicitly in the search statement could potentially be very cumbersome. Therefore, many search engines intended for MeSH indexed data support the so-called explode command [10,13-14,19-20]. When a term is exploded, it is searched along with its narrower terms.

Including narrower concepts in a search (e.g., by issuing an explode command) is sometimes referred to as subsumption. Owing to the principles of indexing with a hierarchic thesaurus, this is an essential retrieval technique. For example, an article on keratitis (inflammation of the cornea) will be indexed with the MesH term “Keratitis” but not with the broader parent term “Corneal Diseases”. A user issuing a query for the MeSH term “Corneal Diseases” without exploding it may not retrieve all articles on narrower concepts such as “Keratitis”.

The MeSH is revised on a yearly basis. Revisions include addition of new terms, deletion or renaming of terms, and changes in the tree structure. This means that databases indexed with MeSH terms should also undergo yearly maintenance (often referred to as “Year-End Processing”) in order to correctly reflect the vocabulary [21]. Not all systems containing MeSH indexing terms have the features required for automation of this process [22]. This means that with time, some of the MeSH terms entered will no longer be consistent with the current version of the MeSH. Also, not all search engines have built-in support for the explode searches required in order to benefit fully from the MeSH tree structure [22].

1.3. The SveMed+ data input system

The data input system ... agement system (RDBMS). The SveM... ation [23]. The MeSH vocabulary is stored in a separate thesaurus database that reflects the current version of the MeSH. The thesaurus database derives from files that are available for download from the NLM[3]. The NLM files are supplemented with translations in Swedish and Norwegian. The database contains information on the vocabulary as such, including data about the MeSH tree structure, and entry term (synonym) vocabulary. The SveMed+ database contains foreign keys (pointers) to the MeSH terms used in the bibliographic references (see Figure 1).

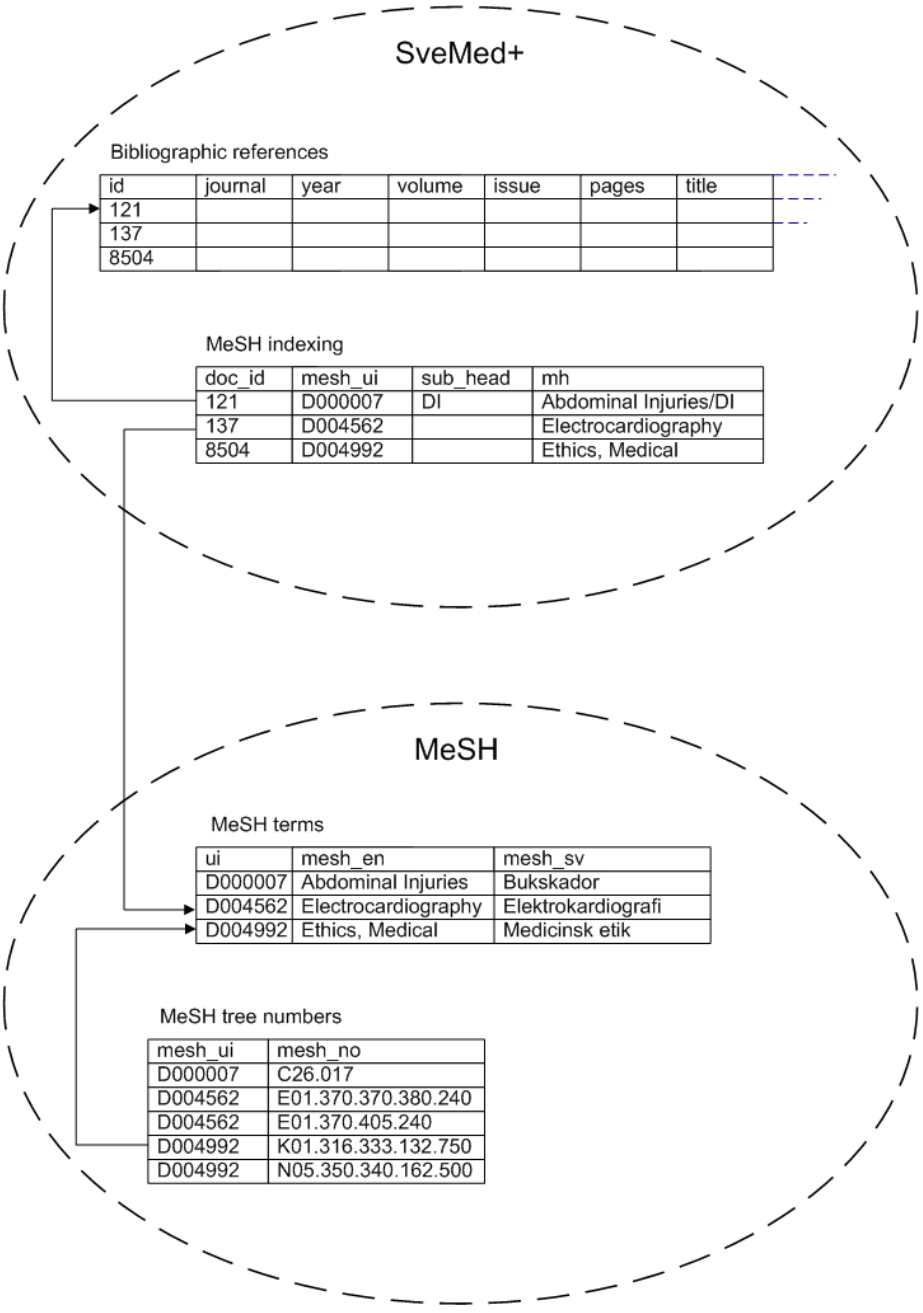

Figure I. The bibliographic references in SveMed+ are associated with MeSH tree numbers via the thesaurus database. Note: This diagram is highly simplified compared to the actual databases.

The indexers have access to a Windows client for data input and maintenance. When indexing, the indexer enters English MeSH terms on[illegible]wedish translations of the terms through the fo[illegible] to base the indexing interface on the Swedish[illegible]their native language. However, the MEDLARS cooperation involved indexing of Scandinavian journals for MEDLINE, allowing the indexers to develop proficiency in the usage of the English version of the thesaurus. Also, the indexers are responsible for the production of the Swedish MeSH translation, which keeps them up-to-date with medical terminology in both languages.

The thesaurus database was initially designed to support maintenance and dissemination of the Swedish MeSH translation produced by KIB. Tables deriving from a Norwegian translation produced by Nasjonalt kunnskapssenter for helsetjenesten in Norway[4] have since been added.

Keeping the thesaurus database up-to-date is essential to the integrity of the SveMed+ database. The thesaurus database is updated on a yearly basis using the files from the NLM. The NLM MeSH files contain unique identifiers that remain the same across different editions of the MeSH. The identifiers are used as keys in the thesaurus database and as foreign keys in the SveMed+ database. Owing to the fact that the keys and foreign keys do not change from one year to another, it is relatively straightforward to adapt the thesaurus database and SveMed+ to new releases of the MeSH.

2. Methods

The data input system described above has served as a data source for several versions of the SveMed+ search interface. In 2009, a project was initiated in order to develop a new version. The project involved a team consisting of developers, web designers and indexers.

Web logs suggest that SveMed+ has a high usage at the Swedish medical schools. It also has a high usage in Norway owing to its coverage of Norwegian medical journals. Judging from the logs and comments received from users, there is a demand for advanced search tools. However, the needs of students and the general public still call for a simple interface.

Users of bibliographic databases typically expect a basic set of retrieval features [24]. The technical specifications for the new version of SveMed+ were based on comments from users and staff as well as the functionality of existing systems for retrieval of bibliographic data in the medical field. The Web interface was implemented using ASP.NET Web Forms with C# as the code behind language and the Visual Studio 2008 integrated development environment as the development tool.

2.1. The Solr index of SveMed+

Although the RDBMS of the data input system is well suited for the data input and maintenance needs of SveMed+, it is not ideal for retrieval of the data. For the purposes of retrieval, the Microsoft SQL server version of SveMed+ is supplemented with a free text index.

A free text index (inverted file) contains words extracted from the documents in the database, arranged alphabetically with pointers to the documents where the words occur [6-7,20,25-28]. In a bibliographic database, the words typically derive from the titles, abstracts, author names and indexing terms of the documents indexed. However, the index may also contain other types of information in order to support specialized features such as chemical searching [29] and explode searches. The free text search engine accesses the index in order to retrieve documents containing the words included in the user query. Retrieval often involves Boolean searching, but other approaches are also possible [25,30].

The free text index of SveMed+ is generated by Solr[5]. Solr is an open source free text search engine based on Lucene [31]. It comes with a data loader and an API upon which a search interface can be built. The API returns hit lists in an XML format (see Figure 2). The SveMed+ interface is based on Solr IR features such as Boolean searching, truncation, stemming, faceting and MoreLikeThis functionality.

In addition to words from the bibliographic records, the Solr index contains tree numbers corresponding to the MeSH terms of the records. Due to the poly-hierarchical structure of the MeSH, each MeSH term may correspond to more than one tree number. For each MeSH term, all the tree numbers in the MeSH hierarchy are loaded into the index. This makes it possible to retrieve records indexed with a MeSH term or its narrower terms by performing a truncation search for the corresponding MeSH numbers. In addition to the MeSH numbers, strings corresponding to MeSH tree numbers with subheadings are stored in the Solr index. There is also a separate field for MeSH terms that are the focus of the article (major headings).

```
<?xml version="1.0" encoding="UTF-8" ?>
- <response>
  - <lst name="responseHeader">
      <int name="status">0</int>
      <int name="QTime">210</int>
    - <lst name="params">
        <str name="facet">true</str>
        <str name="mlt.fl">mesh_en,title_en</str>
        <str name="sort">id desc</str>
        <str name="indent">on</str>
        <str name="facet.mincount">1</str>
        <str name="mlt.mintf">1</str>
        <str name="facet.limit">30</str>
      - <arr name="qf">
          <str>text</str>
          <str>title_en</str>
          <str>title_sv</str>
          <str>mesh_en</str>
          <str>mesh_sv</str>
          <str>tag_en</str>
          <str>mhf</str>
          <str>journal</str>
        </arr>
        <str name="mlt">true</str>
        <str name="defType">edismax</str>
        <str name="version">2.2</str>
        <str name="rows">10</str>
        <str name="mlt.mindf">1</str>
        <str name="start">0</str>
        <str name="q">(((mesh_no:F01.829.500.519*) OR (mesh_no:K01.316*) OR
          (mesh_no:K01.752.256*) OR (mesh_no:N05.350*) OR (etik)) AND
          (tvångsvård))</str>
      - <arr name="facet.field">
          <str>mesh_en</str>
          <str>mesh_sv</str>
          <str>mesh_no</str>
          <str>doctype</str>
          <str>lang</str>
          <str>tag_en</str>
        </arr>
      </lst>
    </lst>
  - <result name="response" numFound="39" start="0">
    - <doc>
      - <arr name="author">
          <str>Ottosson JO</str>
          <str>Ottosson, Jan-Otto</str>
        </arr>
      - <arr name="authorfull">
          <str>Ottosson, Jan-Otto</str>
        </arr>
        <str name="doctype">Artikel</str>
        <str name="email">janotto@bredband2.com</str>
        <str name="firstauthor">Ottosson, Jan-Otto</str>
        <str name="fulltext">Yes</str>
        <int name="id">129009</int>
        <str name="issn">0023-7205</str>
        <str name="journal">Läkartidningen</str>
        <str name="journal_sort">Läkartidningen</str>
        <str name="lang">Swe</str>
      - <arr name="majr">
          <str>Commitment of Mentally Ill</str>
        </arr>
      - <arr name="mesh_en">
          <str>Coercion</str>
          <str>Commitment of Mentally Ill</str>
          <str>Ethics, Medical</str>
          <str>Psychiatry</str>
          <str>Restraint, Physical</str>
          <str>Sweden</str>
        </arr>
      - <arr name="mesh_expsub">
          <str>Y458_N03.706.535.351.200</str>
          <str>Y458.331_F04.096.544.335.200</str>
          <str>Y458.331_N03.706.535.351.200</str>
          <str>Y458_F04.096.544.335.200</str>
        </arr>
      - <arr name="mesh_no">
          <str>B01.050.150.900.649.801.400.112.400.400</str>
          <str>E02.085.700</str>
          <str>E05.472.760</str>
          <str>F04.096.544</str>
          <str>F04.096.544.335.200</str>
          <str>H02.403.690</str>
          <str>I01.880.604.316</str>
          <str>I01.880.630.200</str>
          <str>K01.316.333.132.750</str>
          <str>N03.706.535.351.200</str>
          <str>N05.350.340.162.500</str>
          <str>Z01.542.808.843</str>
        </arr>
      - <arr name="mesh_nor">
          <str>Tvang</str>
          <str>Psykiatrisk tvangsinnleggelse</str>
          <str>Medisinsk etikk</str>
          <str>Psykiatri</str>
          <str>Fysiske tvangsmidler</str>
          <str>Sverige</str>
        </arr>
      - <arr name="mesh_sv">
          <str>Tvång</str>
          <str>Psykiatrisk tvångsvård</str>
          <str>Medicinsk etik</str>
          <str>Psykiatri</str>
          <str>Tvång, fysiskt</str>
          <str>Sverige</str>
        </arr>
        <str name="source">Läkartidningen 2013;110(22)1080-1</str>
      - <arr name="tag_en">
          <str>Humans</str>
        </arr>
      - <arr name="tag_sv">
          <str>Människa</str>
        </arr>
      - <arr name="tag_ui">
          <str>D006801</str>
        </arr>
        <str name="title_en">Ethical aspects of compulsory care</str>
        <str name="title_sv">Etiska aspekter på tvångsvård</str>
        <str name="ui">13077512</str>
        <str
          name="url">http://www.lakartidningen.se/Opinion/Debatt/2013/05/Etiska
          -aspekter-pa-tvangsvard/</str>
        <int name="year">2013</int>
    </doc>
```


Figure 2. The beginning of a Solr hit list. In the SveMed+ free text index, the bibliographic records contain MeSH tree numbers that have been joined in from the thesaurus database. This allows the implementation of an explode feature based on truncation searches for MeSH tree numbers. In this case, the search statement "etik tvångsvård" has been mapped to an explode search for "ethics" (tree numbers F01.829.500.519, K01.316, K01.752.256, and N05.350) that is performed along with a free text search for the words "etik" (ethics) and "tvångsvård" (compulsory care).

The data in the RDBMS is loaded into the free text index on a regular basis. Information about how to extract the words and tree numbers associated with a bibliographic reference is stored in the Solr configuration files. Due to the complex data structure of SveMed+, several SQL JOIN statements have to be performed in order to bring all the data of a bibliographic record together. For each data field, a tokenizer defines how to identify word limits. The MeSH tree numbers require a tokenizer that does not interpret the punctuation marks of the tree numbers as word delimiters.

Although it is the Solr search engine that returns the hit list, many features of the SveMed+ interface rely on the RDBMS for looking up and displaying data.

2.2. Query expansion

Considering that the bibliographic references in databases like MEDLINE and SveMed+ are indexed with terms deriving from the MeSH vocabulary, including MeSH terms in the search is often essential for maximum recall. However, the average user may not be aware of this. Preferably, the search engine should guide the user towards using the MeSH, including features such as explode. Unfortunately, not all search engines come with off-the-shelf support for hierarchical thesauri such as the MeSH.

Query expansion is the process of automatically mapping a search statement (e.g., a natural language query) entered by a user to a query expected to make the most out of the information in the database. For instance, the terms entered by the user may be mapped to terms in the thesaurus [32-36]. Some approaches, such as the Automatic Term Mapping (ATM) feature in PubMed, rely on performing the mapping step behind the scenes [37-44]. Other approaches, such as the Ovid Map Term to Subject Headings feature, rely on presenting the thesaurus to the user in a more interactive way [36,45].

The process of mapping a term to the thesaurus may rely on statistical methods [36] or direct lookups in the thesaurus itself, or possibly a metathesaurus such as the Unified Medical Language System (UMLS) [45-46].

2.3. The Query expansion algorithm of SveMed+

When a user enters a query at the simple search page in SveMed+, the search statement entered is subject to query expansion. The query expansion algorithm was inspired by the PubMed Automatic Term Mapping.

The algorithm was written in the C# programming language. The search string is parsed in order to detect query syntax such as Boolean operators or field tags. The remaining text is mapped against the thesaurus database. If the text contains more than one word, single words as well as multiple word strings are subject to mapping. The string(s) are matched against MeSH terms and their entry terms in English, Swedish and Norwegian. The search string is also mapped against the author field, taking usage of author initials into account. Mapping involves SQL queries against the RDBMS versions of SveMed+ and the MeSH.

The entry term vocabulary of the MeSH consists of synonyms and quasi synonyms of the main headings (preferred terms). This is sometimes referred to as an equivalency relationship. The MeSH thesaurus also has other types of relationships, such as associative ("see related") and hierarchic relationships [12]. It is only the equivalence relationships that come into play in the mapping step. Once mapping has taken place, the MeSH terms (if any) identified by the mapping process are exploded automatically, thereby exploiting their hierarchic relationships with narrower terms.

Since the Solr index contains MeSH tree numbers, an explode search for a MeSH term can be accomplished by performing a truncation search for the corresponding MeSH numbers. A lookup for the numbers to truncate is performed against the thesaurus database under Microsoft SQL Server. All the numbers found must be included in the truncation search in case the corresponding branches of the MeSH tree should contain different sets of narrower terms. Whether or not the search statement was successfully mapped against the MeSH and the author index, a search against the free text index is also performed.

3. Results

The interface of SveMed+ contains a simple search form, an advanced search form and a search form for combining previous search statements from the search history (see Figure 3).

The hit list is displayed along with facets allowing drill-down to more specifics subsets of the search (see Figure 4). The interface also provides functionality such as retrieval of similar documents (see Figure 5) and export to reference management software.

When a user enters a query at the simple search page in SveMed+, it is expanded according to the query expansion algorithm described above. This involves mapping against the MeSH and the author index. The mapping process translates the query into a search statement in the Solr syntax. This search statement will typically contain free text words as well as truncation searches for MeSH tree numbers.

The Solr search returns a hit list and facets. The hits are sorted according to entry date. The facets consist of clusters within the hit list of records sharing a value in the language, document type or MeSH field. The facets are displayed along with the hit list. This allows the user to drill down to a narrower set of records. The query expansion provided by the simple search box is completely transparent to the user and does not require any knowledge of the MeSH thesaurus.

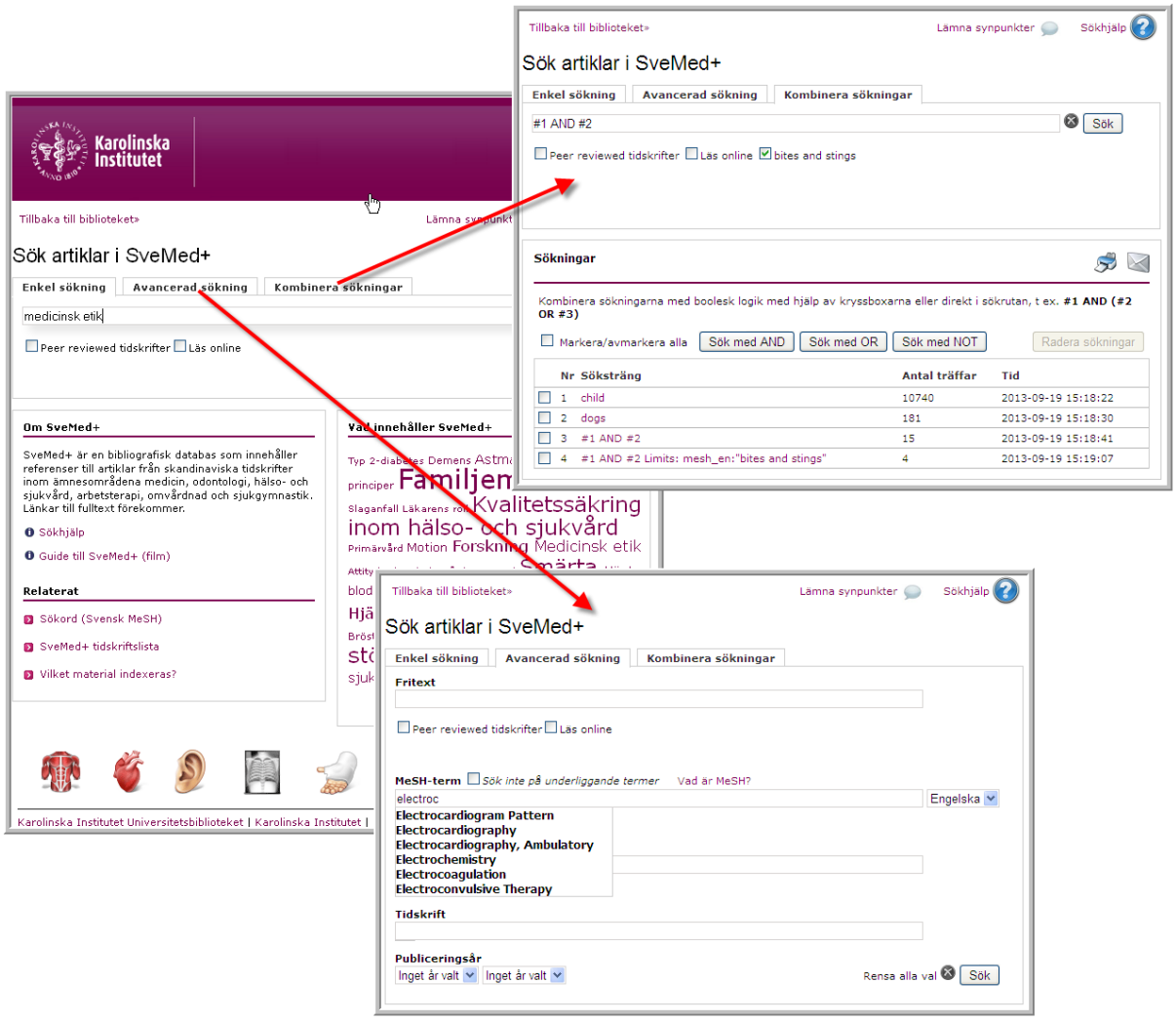

Figure 3. The SveMed+ interface. There are search forms for simple searches, advanced searches, and search history.

Advanced users are offered additional features such as command language searching, support for MeSH subheadings, and index browsing.

The query expansion algorithm is not the only feature in the SveMed+ interface that relies on the MeSH indexing terms. The advanced search page allows the user to browse for a MeSH term in English, Swedish or Norwegian and perform an explode search based on the term selected. The record display in SveMed+ contains a link to similar articles. This feature is based on the Solr MoreLikeThis functionality configured to calculate similarities based on the English MeSH terms and the article title in English. It was inspired by the PubMed Related Articles feature, which is based on similarity according to titles, abstracts and MeSH terms [33]. Finally, some of the facets are MeSH-based. This may draw the attention to potentially useful MeSH terms. There are separate facets for Check Tags and ordinary MeSH terms. This is because the check tags represent indexing terms that are so frequently applied that they might mask relevant terms that are more specific to the search at hand.

Over the years, web logs from SveMed+ have been analyzed on several occasions in order to study various aspects of usage such as overall traffic, distribution of users across different colleges or geographic areas, and usage of individual interface features. Usage of SveMed+ is presently being logged in Google Analytics. During the year 2013, some 315 000 visits were logged. The geographic distribution was 53 % for Norway, 35 % for Sweden and 7 % for Denmark (with insignificant contributions from other countries).

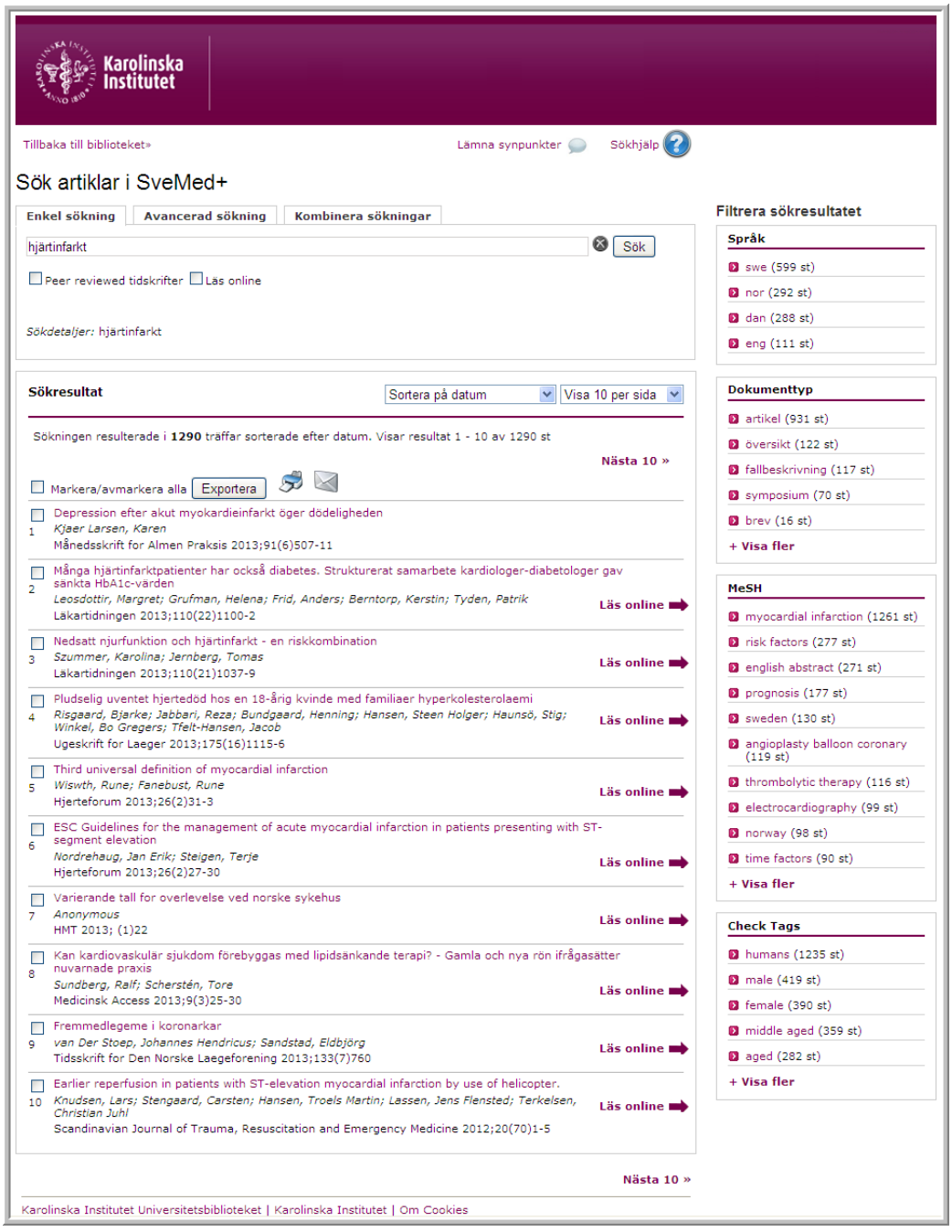

Figure 4. Hit list with facets to the right. The facet "myocardial infarction" represents a MeSH term corresponding to the Swedish term "hjärtinfarkt" entered by the user. The other facets allow the user to drill down to more specific aspects of the subject, such as "risk factors" or "prognosis".

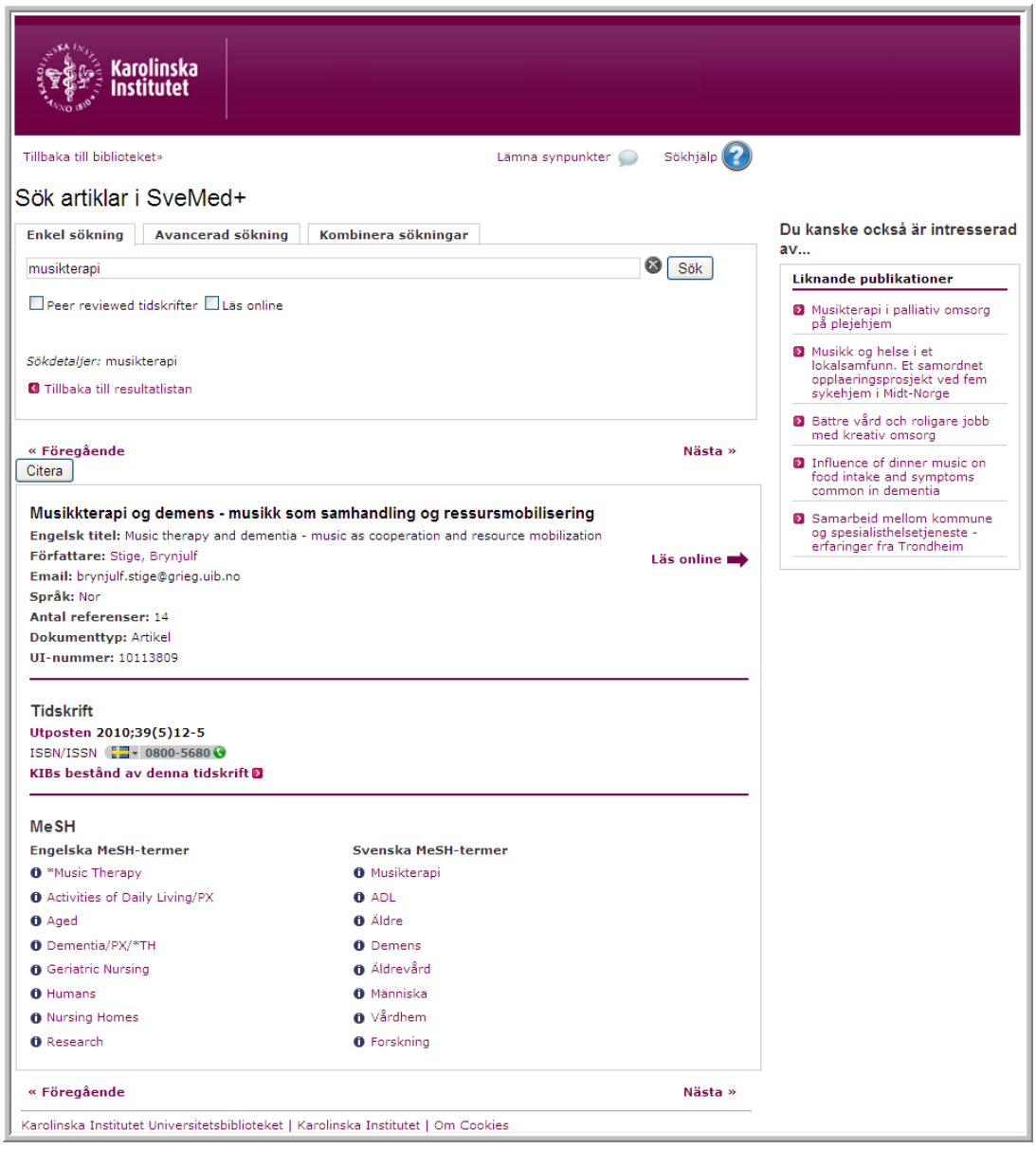

Figure 5. A bibliographic record. The search "musikterapi" has been mapped to The MeSH term "music therapy". Links to similar records are displayed to the right.

Although focusing on Swedish literature, SveMed+ is also of relevance to Danish and Norwegian users since there is no other database with comprehensive coverage of Scandinavian medical journals. In the initial version of the interface, only the English and Swedish versions of the MeSH were used. However, judging from web logs, Norwegian usage of the database is very high. Moreover, the Norwegian user community is very active (in terms of mutual collaboration around SveMed+ and the MeSH, enhancement requests etc). That is why it was it was decided to integrate the Norwegian MeSH translation in the system. The authors of this article are not aware of any Danish MeSH translation.

4. Discussion

A relational database management system (RDBMS) is suitable for data maintenance tasks such as the Year-End-processing of MeSH indexed databases. Retrieval, however, may require indexes tailored for this particular task [47]. The interface of SveMed+ is based on the IR functionality of the Solr search engine as well as the relational database (RDBMS) behind the data input system.

The simple search box of SveMed+ relies on query expansion in order to improve recall. In similarity to the PubMed Automatic Term Mapping (ATM), the query expansion algorithm maps the terms entered by the user to MeSH terms that are automatically exploded. However, unlike the ATM, it also performs multilingual mapping based on the Swedish and Norwegian translations of the MeSH. The fact that the database contains article titles in several languages and their translations in English contributes to some additional multilinguality.

The implementation of multilinguality in search interfaces may involve problems associated with homonymy (i.e., the same word meaning different things in different languages). For example, the Swedish MeSH term "Glass" means ice cream whereas the Norwegian MeSH term "Glass" means glass. In the advanced search interface, the user can select language when browsing the MeSH index, thereby eliminating any ambiguity caused by homonymy between languages. However, there is no way for the user to indicate which language is used when a query is entered in the simple search box. This causes some queries (such as "glass") to produce unexpected results. Fortunately, the MeSH translations used in the database only contain a limited number of homonyms (in most cases associated with closely related terms, such as "Elektrostimulering" for electric stimulation and electric stimulation therapy, respectively). However, with a very large thesaurus containing many terms and languages, the present implementation of multilinguality would possibly be a bit problematic.

The searches generated through query expansion are performed against a Solr based free text index, but the query expansion algorithm also relies on lookups against the RDBMS in order to map user input against MeSH terms and author names. MeSH mapping typically results in a Solr search containing a mixture of free text words and truncated MeSH tree numbers. Some of the IR features of the Solr search engine, such as relevancy ranking, cannot be applied with this approach. However, features such as facets, stemming and MoreLikeThis can still be used.

Although the query expansion is something that happens automatically, the user is alerted to MeSH terms in the records retrieved through the facets provided by Solr. Automatic query expansion and stemming may improve the recall of searches performed by inexperienced users. However, advanced users may be a bit confused when the search engine adjusts the query behind the scenes. In particular, truncation searches may not work the way the user expects.

The query expansion involves performing automatic explode searches based on the MeSH hierarchy. Search engines such as BRS/Search and Oracle Text have off-the-shelf support for hierarchical searches. In the present version of SveMed+, explode searches are implemented through tree numbers stored in an inverted file. The ELLHILL system had a similar approach [48]. NCBI Entrez, the retrieval system behind PubMed, relies on loading the postings for each MeSH term and MeSH/subheading combination in a separate inverted file (NCBI staff, personal communication).

In Svemed+, searches for narrower terms are accomplished by truncation searches for tree numbers in the Solr index. Penn Museum has developed an interface where searches for narrower and alternate terms are implemented through the Solr SynonymFilerFactory [49]. In this case, the parent terms of each term appearing in the controlled vocabulary are loaded as synonyms.

Taking the thesaurus hierarchy into account when generating the inverted file (or files) is not the only possible approach when implementing explode searches. Associations between a query term and its narrower terms could also be made entirely on-the-fly. For example, the task of finding the tree numbers of narrower terms and retrieving the corresponding records can be accomplished through SQL statements generated at query time [16,50]. Performing some of the logic associated with explode (truncations, SQL joins etc) when generating the inverted file instead of at query time may offer a performance advantage. The choice of approach should be guided by the database size and the intended application.

Although the information about which MeSH terms have been added to the bibliographic records is stored in the SveMed+ database, thesaurus features such as MeSH mapping and explode searches rely on a separate thesaurus database. The thesaurus database is multilingual. Initially developed for maintenance of the Swedish MeSH translation, it has subsequently been supplemented with the Norwegian MeSH translation as a service to the Norwegian users of SveMed+. The mapping process could easily be extended to include other MeSH translations or vocabularies (e.g., by using the UMLS).

In summary, bibliographic databases may call for rather sophisticated retrieval functionality, such as support for hierarchical thesauri. Developers of interfaces to bibliographic systems need to be aware of these user needs. Not all software packages come with off-the-shelf support for the features expected by the users. By combining various tools

and approaches, it may still be possible to meet user expectations. SveMed+ provides an example of how this can be done in practice.

Notes

1. http://svemedplus.kib.ki.se
2. http://mesh.kib.ki.se
3. http://www.nlm.nih.gov/mesh/
4. http://www.kunnskapssenteret.no/
5. http://lucene.apache.org/solr/

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Haglund L. 200 years of medical information - some landmarks. Journal of the European Association for Health Information and Libraries 2010; 6(4): 21-28.
[2] Almling M. MIC - a glimpse of the past. Stockholm: Karolinska Institute Library and Information Center, 1992.
[3] Knutssön G and Monik M. News from the special interest group on mesh. Journal of the European Association for Health Information and Libraries 2010; 6(2): 48-50.
[4] Corning ME. The U.S. National Library of Medicine and international MEDLARS cooperation. Information Storage and Retrieval 1972; 8: 255-264.
[5] Kenton C. MEDLINE searching and retrieval. Medical Informatics 1978; 3(3): 225-235.
[6] Beckelhimer MA, Cox JW, Hutchins JW and Kenton DL. The MEDLINE hardware and software. Medical Informatics 1978; 3(3): 197-209.
[7] McCarn DB. MEDLINE: An introduction to on-line searching. Journal of the American Society for Information Science 1980; 31(3): 181-192.
[8] Dee CR. The development of the medical literature analysis and retrieval system (MEDLARS). Journal of the Medical Library Association 2007; 95(4): 416-425.
[9] Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association 2000; 88(3): 265-266.
[10] Sewell W. Medical subject headings in MEDLARS. Bulletin of the Medical Library Association 1964; 52(1): 164-170.
[11] Cain AM. Thesaural problems in an on-line system. Bulletin of the Medical Library Association 1969; 57(3): 250-259.
[12] Nelson SJ, Johnston D and Humphreys BL. Relationships in medical subject headings (MeSH). In: Bean CA and Green R (eds) Relationships in the organization of knowledge. Dordrecht: Kluwer Academic Publishers, 2001, pp. 171-184.
[13] Coletti MH and Bleich HL. Medical subject headings used to search the biomedical literature. Journal of the American Medical Informatics Association 2001; 8(6): 317-323.
[14] Lowe HJ and Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 1994; 271(14): 1103-1108.
[15] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy J-P, et al. Enhancing the MeSH thesaurus to retrieve french online health resources in a quality-controlled gateway. Health Information and Libraries Journal 2004; 21: 253-261.
[16] Gavel Y, Andersson P-O and Knutssön G. Euroethics - a database network on biomedical ethics. Health Information and Libraries Journal 2006; 23(3): 169-178.
[17] Nelson SJ, Schopen M, Savage AG, Schulman J-L and Arluk N. The MeSH translation maintenance system: Structure, interface design, and implementation. MEDINFO 2004; 11(1): 67-69.
[18] Liu F, Fontelo P and Ackerman M. BabelMeSH: Development of a cross-language tool for MEDLINE/PubMed. AMIA Annual Symposium proceedings 2006: 1012.
[19] Hersh WR. Retrieval. In: Information retrieval : A health and biomedical perspective. New York: Springer, 2009, pp. 199-233.
[20] Kellerman FR. Computerized bibliographic searching: MEDLINE and beyond. In: Introduction to health sciences librarianship : A management handbook. Westport, Conn.: Greenwood Press, 1997, pp. 57-84.
[21] Humphrey SM. File maintenance of MeSH headings in MEDLINE. Journal of the American Society for Information Science 1984; 35(1): 34-44.
[22] McGregor B. Medical indexing outside the National Library of Medicine. Journal of the Medical Library Association 2002; 90(3): 339-341.
[23] Date CJ. An introduction to database systems. 6th edition. Reading, Mass.: Addison-Wesley Publishing Company, 1995.
[24] Othman R and Halim NS. Retrieval features for online databases: Common, unique, and expected. Online Information Review 2004; 28(3): 200-210.
[25] Zobel J and Moffat A. Inverted files for text search engines. ACM Computing Surveys 2006; 38(2).

[26] Harman D, Fox E, Baeza-Yates R and Lee W. Inverted files. In: Frakes WB and Baeza-Yates R (eds) Information retrieval : Data structures & algorithms. Englewood Cliffs, N.J: Prentice Hall, 1992, pp. 28-43.
[27] Hersh WR. Indexing. In: Information retrieval : A health and biomedical perspective. New York: Springer, 2009, pp. 159-197.
[28] Doszkocs TE. From research to application: The CITE natural language information retrieval system. In: Salton G and Schneider H-J (eds) Research and development in information retrieval. Germany: Springer-Verlag, 1983, pp. 251-262.
[29] Schultheisz RJ, Walker DF and Kannan KL. Design and implementation of an on-line chemical dictionary (CHEMLINE). Journal of the American Society for Information Science 1978; 29(4): 173-179.
[30] Wiesman F, Hasman A and van den Herik HJ. Information retrieval: An overview of system characteristics. International Journal of Medical Informatics 1997; 47: 5-26.
[31] Smiley D and Pugh E. Solr 1.4 enterprise search server : Enhance your search with faceted navigation, result highlighting, fuzzy queries, ranked scoring, and more. Birmingham, UK: Packt Publishing, 2009.
[32] Shiri AA, Revie C and Chowdhury G. Thesaurus-enhanced search interfaces. Journal of Information Science 2002; 28(2): 111-122.
[33] Lin J and Wilbur WJ. PubMed related articles: A probalistic topic-based model for content similarity. BMC Bioinformatics 2007; 8: 423.
[34] Srinivasan P. Query expansion and MEDLINE. Information Processing & Management 1996; 32(4): 431-443.
[35] Schoonbaert D. Automatic mapping of free-text to thesaurus: Good policy? Bulletin of the Medical Library Association 1997; 85(4): 439-440.
[36] Shiri A and Revie C. Query expansion behaviour within a thesaurus-enhanced search environment: A user-centered evaluation. Journal of the American Society for Information Science and Technology 2006; 57(4): 462-478.
[37] NLM. PubMed tutorial - automatic term mapping, http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html (2013, accessed September 2013)
[38] NLM. How PubMed works: Automatic term mapping, http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.How_PubMed_works_aut (2013, accessed September 2013)
[39] Knecht LWS and Nelson SJ. Mapping in PubMed. Bulletin of the Medical Library Association 2002; 90(4): 475.
[40] Nahin AM. Boost for PubMed search results: New atm & citation sensor introduced. NLM Technical Bulletin 2008; (362): e10.
[41] Griffon N, Chebil W, Rollin L, Kerdelhue G and Thirion B. Performance evaluation of a unified medical language system's synonym expansion to query PubMed. BMC Medical Informatics and Decision Making 2012; 12: 12.
[42] Thirion B, Robu I and Darmoni SJ. Optimization of the PubMed automatic term mapping. In: Adlassnig K-P (ed) Medical informatics in a united and healthyEeurope: Proceedings of mie 2009, the xxii international congress of the European federation for medical informatics. Amsterdam: IOS Press, 2009.
[43] Lu Z, Kim W and Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. Information Retrieval 2009; 12: 69-80.
[44] Canese K, Jentsch J and Myers C. PubMed: The bibliographic database. In: McEntyre J and Ostell J (eds) The NCBI handbook. Bethesda (MD): National Center for Biotechnology Information, 2002.
[45] Gault LV, Shultz M and Davies KJ. Variations in medical subject headings (MeSH) mapping: From the natural language of patron terms to the controlled vocabulary mapped lists. Journal of the Medical Library Association 2002; 90(2): 173-180.
[46] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research 2004; 32(Database issue): D267-D270.
[47] Ostell J. Databases of discovery. ACM queue 2005; 3(3): 40-48.
[48] MEDLARS II staff operating manuals. Santa Monica: System Development Corporation, 1973-1974.
[49] Williams S. Better search through query expansion using controlled vocabularies and Apache Solr. Code4Lib [Internet]. 2013; (20), http://journal.code4lib.org/articles/7787 (2013, accessed January 2014)
[50] Oliver DE, Bhalotia G, Schwartz AS, Altman RB and Hearst MA. Tools for loading MEDLINE into a local relational database. BMC Bioinformatics 2004; 5(146).