ELSEVIER

Using CISMeF MeSH ''Encapsulated'' terminology and a categorization algorithm for health resources

Aurélie Névéol [a,b], Lina F. Soualmia [a,b], Magaly Douyère [b], Alexandrina Rogozan [a], Benoît Thirion [b], Stefan J. Darmoni [a,b,*]

[a] *Perception and Information Systems Laboratory CNRS FRE-2645 & L@STICS, Medical School, University of Rouen, Rouen Cedex, France*
[b] *CISMeF, Rouen University Hospital, Rouen Cedex, France*

Summary *Introduction:* CISMeF is a Quality Controlled Health Gateway using a terminology based on the Medical Subject Headings (MeSH) thesaurus that displays medical specialties (metaterms) and the relationships existing between them and MeSH terms. *Objective:* The need to classify the resources within the catalogue has led us to combine this type of semantic information with domain expert knowledge for health resources categorization purposes. *Material and methods:* A two-step categorization process consisting of mapping resource keywords to CISMeF metaterms and ranking metaterms by decreasing coverage in the resource has been developed. We evaluate this algorithm on a random set of 123 resources extracted from the CISMeF catalogue. Our gold standard for this evaluation is the manual classification provided by a domain expert, viz. a librarian of the team. *Results:* The CISMeF algorithm shows 81% precision and 93% recall, and 62% of the resources were assigned a ''fully relevant'' or ''fairly relevant'' categorization according to strict standards. *Discussion:* A thorough analysis of the results has enabled us to find gaps in the knowledge modeling of the CISMeF terminology. The necessary adjustments having been made, the algorithm is currently used in CISMeF for resource categorization.

1. Introduction

The Internet has become a very prosperous source of information in numerous fields, including health. Users are now experiencing huge difficulties in finding precisely what they are looking for, among the tons of documents available online. Generic search engines cannot solve this problem efficiently because they usually offer a selection of documents that turns out to be either too large or ill suited to the query. In this context, the CISMeF project (French acronym of Catalogue and Index of Health Resources in French) was initiated in 1995. CISMeF is a Quality Controlled Health Gateway [1] cataloguing the most important resources of institutional health information in French, and it is manually maintained. In CISMeF, a resource is defined as (1) a Web site or (2) high-quality documents from this Web site. CISMeF describes and indexes the most important sources of institutional health

*Corresponding author. Tel.: +33-2-32-88-88-29; fax: +33-2-32-88-88-32.
E-mail address: stefan.darmoni@chu-rouen.fr (S.J. Darmoni).

information in French, in order to allow one to search them quickly and precisely. A great variety of resources is indexed, in terms of resource type (medical guidelines, health law documents, course material, information for patients, etc.), and resource format (html, pdf, etc.). The catalogue currently contains 13,100 resources and is consulted by 20,000 visitors per day. The catalogue is steadily expanding with 55 new resources per week. Adding a new resource to the catalogue is a four-step process. First, a daily thorough technological scanning on the Internet retrieves new resources. Then, some of these resources are selected according to specific quality criteria based on the NetScoring policy (URL: http://www.chu-rouen.fr/netscoring) and on the European Union-funded MedCIRCLE project (URL: http://www.medcircle.info). The resources that have been selected are indexed manually, and finally included in the catalogue. Indexing is a decisive step for the efficiency of information retrieval within CISMeF.

2. Objectives

This experiment aims to provide an automatic categorization system for CISMeF resources. The categorization is designed to provide the reader with an enhanced resource description that will allow a quick and easy assessment of the main topics discussed in the resource. In CISMeF, this type of categorization will list the medical specialties relevant to a resource by decreasing order of their importance in the text.

Since 1995, CISMeF has always been evolving towards this direction. First, the documents were indexed with medical keywords (e.g. *AIDS*, *diabetes*) that were coordinated with qualifiers (e.g. *diagnosis*, *epidemiology*) if necessary. Qualifiers are used to better define a topic, or express a certain aspect of a keyword, e.g. *AIDS/diagnosis*. We will now refer to keywords and qualifiers as ''terms''. Then, the concept of ''major'' and ''minor'' keywords (or keyword/qualifier pair) has been introduced. A term is said to be major if the concept it represents is discussed throughout the whole document. It is said to be ''minor'' if it is referred to only in a few paragraphs. Major terms are marked in CISMeF by a star. Since the number of keywords indexing a document can be quite large (a few dozens)—especially if the resource is a clinical guideline—and since these keywords usually refer to narrow and precise concepts, it is necessary to establish a different kind of classification, one that would briefly inform the user about the specific medical fields and specialties covered in the document. These medi-

cal specialties are inferred from the existing CISMeF terminology. They used to be sorted in alphabetical order in the first place. This work aims to test the relevance of a new algorithm inducing medical specialties and ranking them by decreasing coverage in the resource. The algorithm ranking is compared to the manual categorization performed by a medical librarian, which is considered as our gold standard. As pointed out by Bodenreider [2] in a similar work, using such a categorization method may also provide an evaluation of the terminology.

2.1. Background: CISMeF MeSH 'encapsulated' terminology

To understand the idea of categorization within CISMeF, it is important to have a clear idea of how the CISMeF terminology is structured [3,4]. CISMeF uses the Medical Subject Headings (MeSH) thesaurus from the US National Library of Medicine [5]. The MeSH thesaurus contains about 22,000 MeSH keywords and 84 qualifiers in its year 2003 version.

CISMeF resource description has two levels.

The first level is composed of two sets of information:

- a list of MeSH keywords, coordinated (or not) with MeSH qualifiers;
- a list of resource types [6] (a resource type is an indication on the nature of the document, e.g. teaching material). The CISMeF resource types are an extension of Medline publication types. They were introduced to cope with the heterogeneity of Internet health resources. Currently, 128 different resource types are discerned. The list of resource types is available at the following URL: http://www.chu-rouen.fr/documed typeeng.html.

These three types of information (MeSH keywords, MeSH qualifiers and resource types) are structured according to the MeSH hierarchy for keywords and qualifiers, and to the CISMeF resource types hierarchy.

The second level consists of a list of metaterms [4]. A metaterm is generally a medical specialty or a biological science (e.g. *cardiology* or *bacteriology*) selected by the CISMeF chief librarian. There are 67 different metaterms in CISMeF. For each metaterm, semantic links were created with one or more MeSH keywords, qualifiers and resource types. For example, the metaterm *psychiatry* is associated with the MeSH keywords *psychiatry* and *psychiatric hospital* that belongs to a completely different tree structure within the MeSH and also with the CISMeF resource type *mental health dispensary*.

In fact, the idea of creating metaterms came up to optimize information retrieval in CISMeF and to cope with the relatively restrictive nature of MeSH keywords. For instance, the queries 'guidelines in cardiology' and 'databases in psychiatry' where *cardiology* and *psychiatry* are only MeSH keywords get few or no answers. Introducing *cardiology* and *psychiatry* as metaterms is an efficient strategy to get more results because instead of exploding one single MeSH tree (e.g. *psychiatry* as a MeSH keyword), the use of metaterms results in automatic expansion of the queries by exploding other related MeSH or CISMeF trees in addition to the current tree (e.g. *psychiatric hospital* as a MeSH keyword or *mental health dispensary* as a resource type will be exploded in the case of the *psychiatry* query). The list of metaterms is available at the following URL: http://www.chu-rouen.fr/ssf/santspeeng.html.

3. Material and methods

3.1. CISMeF categorization algorithm

The categorization algorithm presented here is based on the CISMeF librarians' technical know-how, and exploits the resource indexing at our disposal in the CISMeF database. The categorization algorithm uses all the semantic links existing between MeSH keywords, qualifiers and resource types of a resource indexed in CISMeF and metaterms to induce the list of metaterms for that resource. As domain experts, CISMeF librarians defined a scoring procedure to assign a score to each metaterm, and rank the list. If a MeSH keyword has a link to several metaterms, it can induce more than one metaterm. For example, the keyword *thumb* induces the metaterm *anatomy*, and the keyword *alcoholism* induces both the metaterms *psychiatry* and *toxicology*. Similarly, the MeSH keyword/qualifier pair *alcoholism/legislation and jurisprudence* induces the metaterms *psychiatry* (from the semantic link with *alcoholism*), *toxicology* (from the semantic link with *alcoholism*) and *medical law* (from the semantic link with *legislation and jurisprudence*). The resource type *mental health dispensary* induces the metaterm *psychiatry*, etc.

Assume there are n MeSH terms M_1, M_2^*, ..., M_n (major terms are marked by a star), m qualifiers Q_1, Q_2^*, ..., Q_m (majors qualifiers coming from major keyword/qualifier pairs) and p resource types R_1, ..., R_p assigned to a resource. The CISMeF terminology enables us to infer k metaterms T_1, T_2, ..., T_k from these sets of terms. For each metaterm, a major score and a minor score are computed:

- major(T_i) = Card{M^*, Q^*, R, inducing T_i}, the number of *major* indexing terms and resource types which induce T_i
- minor(T_i) = Card{M, Q, inducing T_i}, the number of *minor* indexing terms which induce T_i

Hence, metaterms are classified by decreasing major scores, and in the case of similar major scores, minor scores are used to obtain the final semantic categorization.

Metaterms with major scores higher than zero are said to be major metaterms, and they are represented with a number of stars corresponding to their major scores. For example, if the metaterm *psychiatry* is induced from the major MeSH keyword *psychiatric hospital*, the minor MeSH keyword *psychiatry* and the resource type *mental health dispensary* it will be awarded two stars. As an illustration, Fig. 1 shows the categorization ('specialties') obtained with the CISMeF algorithm for a sample resource of the CISMeF catalogue.

3.2. Evaluation

To test the relevance of the categorization method presented above, the automatic categorization obtained was compared to the classified list of metaterms (or medical specialties) provided by a CISMeF librarian for each resource. The whole process of manually assigning metaterms, running the CISMeF algorithm on the manual indexing previously performed, and evaluating the results requires an average of 20 min per document. Therefore, the evaluation of the CISMeF categorization method has been performed on a sample of 123 resources randomly selected in the CISMeF database. Table 1 gives a breakdown of the number of resources according to the number of metaterms assigned by the librarian.

Manual categorization is considered as our gold standard (i.e. it is assumed to provide a fully relevant list of properly ranked specialties) and was

Table 1 Number of metaterms (medical specialties) to extract from the resources

Number of specialties	Number of resources	%
1 Specialty at most	31	25.20
2 Different specialties	32	26.02
3 Different specialties	32	26.02
4 Or more specialties	28	22.76
Total	123	

Notice Doc'CISMeF

Title : *Assistance to HIV-AIDS positive patients*
Sous-Title : A survey conducted with care units patients in the Paris area and the
Provence Alpes Côte d'Azur region. *Une enquête auprès des patients des services de soins en Ile
de France et en Provence - Alpes - Côtes d'Azur*

CATEGORIZATION

Specialties :   **** allergy and immunology
                    **** virology
                    ** epidemiology
                    ** statistics
                     environment and public health
                    economics psychiatry
                    toxicologie toxicology

Keywords :   sex distribution
                 socioeconomic factors
                 FRANCE
                 * self-help groups
                 *HIV infections
                 *HIV infections / epidemiology
                 HIV infections / psychology
                  housing
                 physician's role
                 *acquired immunodeficiency syndrome
                 * acquired immunodeficiency syndrome / epidemiology
                 acquired immunodeficiency syndrome / psychology
                 * social support
                 substance-related disorders

Type(s) :   journal article

IDENTIFICATION

URL(s) :   http://www.sante.gouv.fr/drees/etude-resultat/er-pdf/er203.pdf

Fig. 1 Categorization of the resource entitled 'Assistance to HIV-AIDS positive patients' (Aides apportées aux personnes atteintes par l'infection à VIH-SIDA) http://www.sante.gouv.fr/drees/etude-resultat/er-pdf/er203.pdf.

produced after the librarian saw the results of the automatic categorization: the evaluator was not blind to the results. The librarian proof-read the automatic categorization, in the sense that, after a thorough inspection of the resource, every metaterm retrieved by the algorithm was approved or discarded, the minor/major weight allotted to the approved metaterms was checked, non retrieved metaterms were added to the list if necessary, and the ranking was also checked. The result of this adjustment work was considered as the proper set of metaterms.

The algorithm was evaluated at two different levels: first, at the level of the metaterms extracted (using precision and recall) and at the resource level (using a relevancy scale). Precision is the ratio between the number of metaterms correctly extracted by the algorithm and the total number of extracted metaterms. Recall is the ratio between the number of metaterms correctly extracted by the algorithm and the total number of relevant metaterms for the collection of resources analyzed. We also give values for silence (1-recall) and noise (1-precision), which are the usual measures in information science.

The evaluation was designed to rate the relevance of the specialties extracted by the algorithm, according to a Likert Scale described below: ''Fully Relevant'' means that manual and automatic categorizations are exactly similar, ''Partially Relevant'' means that the automatic categorization is quite similar to the manual categorization (more specifically, ''fairly relevant'' means there are two minor errors at most, or less than 50% error, ''average relevance'' means there are three or four errors, or 50% error and ''poorly relevant'' means there are more than four errors, or more than 50% error). ''Non-Relevant'' means that the automatic classification is absolutely not similar to the manual categorization (100% error), or that there is a silence (even minimal) of the automatic categorization.

4. Results

The results show 80.8% precision (overall, 298 out of the 369 specialties selected by the algorithm were relevant)—which corresponds to 19.2% noise, and 93.4% recall (overall, the algorithm selected 298 specialties out of the 319 that were expected)—

Table 2   Overall categorization process

Relevance	Number of Resources	%
Fully relevant	45	36.58
Partially relevant (good)	32	26.01
Partially relevant (average)	20	16.26
Partially relevant (bad)	5	4.06
Non-relevant	21	17.07

Table 3   Categorization errors

Error type	Number of resources
Inappropriate specialty (noise)	36
Missing specialty (silence)	21
Ranking error	39
Weighting error	
Major instead of minor	42
Minor instead of major	7

which corresponds to 6.6% silence. These results show that the automatic categorization of resources is nearly exhaustive (less than 7% of silence) and also quite reliable (less than 20% of noise).

An indication of the relevance of the categorization for each resource is presented in Table 2. Two third (62.6%) of the resources are assigned a ''fully relevant'' or ''fairly relevant'' categorization, while 17.1% were assigned a ''non relevant'' categorization, mainly as a consequence of silence. Table 3 shows the occurrences of each error type described in the previous section. The most frequent errors are ranking and weighting errors, which means that the appropriate specialties are usually retrieved.

5. Discussion

5.1. Evaluation results

The main results of this work (80.8% precision and 93.4% recall) are very encouraging, especially considering that the terminology features used by the algorithm, namely, the metaterms, were intended to enhance Information Retrieval in the first place. Based on these results we decided to categorize all 13,100 resources in the CISMeF catalogue with the algorithm described above.

The algorithm produces the exact (i.e. recommended by the librarian) categorization for more than one third (36.6%) of the resources. The cases

where the categorization is considered as ''non relevant'' result from resources for which some silence is involved. The overall silence of 7% results from 21 specialties out of the 319 expected ones that could not be selected by the algorithm. This silence has an important impact on the categorization of the resources—at least on the resources in the sample test set, as it causes 21 documents out of 123 (i.e. 17.1%) to be assigned a ''non relevant'' categorization. In many of the ''partially relevant'' cases, the errors are light. In fact, very few cases combine all the error types. Although we did not conduct a blind evaluation, this method enabled the librarian to assess the relevance of the automatic categorization, so as to focus on the actual noise and silence of the algorithm.

The silence of the categorization is mainly restricted to specific domains in the terminology. For example, a resource that was indexed with the keywords *travel* and *tropical medicine*, and the resource type *popular works* can not be properly categorized by the algorithm because there is no link between these keywords and any metaterm. Hence, analysing the results led us to uncover several similar cases, where links to existing metaterms have to be created in the terminology, or new metaterms have to be created along with the relevant links. In the previous example, *tropical medicine* is a medical specialty and it is thus necessary to create a metaterm for *tropical medicine*, which will be linked to the keyword *tropical medicine*. Such metaterm creations will result in a better coverage of the CISMeF terminology while enhancing the performance of the categorization algorithm. Moreover, it may also enhance information retrieval within CISMeF by providing more answers to the queries.

In spite of these lacks in the CISMeF terminology, which have been made up for, the overall silence is low, probably because these lacks would mainly affect specialties that were little covered in CISMeF, or at least in the evaluation sample (e.g. *biochemistry*, *cellular biology* or *plastic surgery*).

However, it is impossible to reduce the noise without cutting the performances of information retrieval. In fact, reducing the noise would imply identifying the semantic links causing it and suppressing them. It is also important to mention that most of the metaterms that the algorithm produces, although the librarian considers them unnecessary for a proper categorization, are not irrelevant to the document *per se*, they simply make the categorization a little broader than intended.

Regarding the major/minor weighting of the terms, the algorithm mistakes a minor term for a major term (42 errors) six times more than it

mistakes a major term for a minor one (7 errors). This most certainly results from our mislead assumption that resource types, which are not given any weight in the first place, should be considered as ''major'' terms for categorization purposes. Therefore, a major/minor weighting scheme for resource types will be introduced in the CISMeF terminology, so that resource types will be processed in the same way as keywords and qualifiers by the categorization algorithm.

5.2. Literature review

There are different approaches to automatic text categorization: linguistic, semantic and machine learning approaches. A linguistic approach exploits a formal representation of the grammar of the language in which the document is written [7]. A semantic approach exploits relations between medical concepts and does not necessarily require a linguistic component. There are a number of machine learning methods: naive Bayes, k-nearest neighbors (k-NN), decision trees, neural networks, support vector machines (SVM), and statistical compression models [8].

Our work is positioned in the semantic approach as it exploits a formal representation of medical concepts. The CISMeF algorithm does not use any natural language processing.

Prior to text categorization, several automatic indexing projects were initiated [8—11]. The most acknowledged project is probably the US NLM Indexing Initiative (IND) [9] which mixes the three approaches (linguistic, semantic and machine learning) and includes the MetaMap program [10]. Automatic indexing produced by IND compare very favorably with the standard human indexing (no statistical difference) [8].

Bodenreider [2] then conducted an automatic text categorization study similar to the one presented here, using the Unified Medical Language System (UMLS) metathesaurus semantics, after a mapping of index terms to selected keywords. It is not possible to use the UMLS in French at the moment, because there is no complete French version available yet. However, the VUMeF consortium is currently working on this issue [12].

In recent work, Teehan and Harper [13] show that statistical compression models, and in particular PPM (Prediction by Partial Match) models, have interesting performances when used for text categorization of newspaper articles. Unlike most techniques, this type of global approach does not require keyword extraction prior to categorization.

Another novel approach to text categorization is that of SVM [14], which seem to perform very well, even in multi-class problems [15], up to about 20 classes. A study of k-NN classifiers on histology reports classification concluded that k-NN was neither a very reliable nor efficient classification method.

Other strategies consist in combining both statistical and linguistic approaches. For instance, Mehta et al. [16] and Han and Kamber [17] show how text categorization can be treated as a text database mining problem. Wilcox and Hripcsak [18] also uses both linguistic and statistical methods to determine the presence of six medical conditions in chest cardiograph reports. The classical analytical methods used for text categorization are either inductive learning methods such as decision trees, Bayesian classifiers and Bayesian Networks, or rule-based methods. The results show that rule-based methods (or combined methods including a rule-based technique) give the best results, provided that an accurate modeling of the categorization process and domain concepts can be obtained from domain experts.

5.3. Comparison with other research

In this context, we have decided to implement a scoring algorithm, based on the CISMeF terminology, and on the technical know how of the librarians in this team, who are experts in medical document indexing and categorization. Compared to the other analytical approaches quoted above, our algorithm does not need training. This makes the algorithm easily applicable, but on the other hand there are no automatic means to improve its performance over time. The results of this study enabled us to make several manual changes to both the knowledge modeling and scoring procedure in order to improve the categorization process.

How do our results compare to other similar work in the medical domain?

SVM and PPM compression have not been used yet for the categorization of medical resources. The results of the UMLS study [2] show that in 90% of the cases, the categorization is ''fully relevant'' (i.e. no noise, no silence), compared to 61% with our algorithm. We performed categorization among 67 different specialties, whereas Bodenreider [2] was working with 22 disease categories only.

We obtained better overall results than the combined classifiers tested by Larkey and Croft [19], who concluded that their system should be used as an interactive help to human indexers rather than an unsupervised classifier. In spite of the noise and silence, our results show that the CISMeF algorithm is able to meet high standards in terms of ranking and major/minor weighting in two third of the cases.

On a more general scale, the CISMeF categorization algorithm could be used to categorize health resources indexed by English MeSH keywords, as all the CISMeF terminology, including the adds-on (metaterms and resource types) is entirely bilingual. (French and English). Therefore, the method presented above could be tested for Internet health resources categorization in prominent health gateways such as OMNI (URL: http://omni.ac.uk/), Cliniweb (URL: http://www.ohsu.edu/cliniweb/) or HealthInSite (URL: http://www.healthinsite.gov.au/). Besides Internet resources, the CISMeF categorization algorithm can also be used to categorize scientific articles, and in particular those of the Medline bibliographic database. In this case, the resource types should be restricted to Medline publication types. The categorization of articles from this database, or from scientific journals (*IJMI*, *Nature* or *Science*) would characterize their contents by bringing out the medical specialties covered by each source, etc.

This categorization seems also valid for electronic patient records. In such a situation, the task is recognized as difficult due to inequalities in the precision and style of the text. The CISMeF categorization should be adapted to the users broad implicit categorization, which means no more than one hundred categories. The categorization algorithm could be used on electronic records previously indexed with MeSH terms (the indexing being either manual or automatic). The resulting categorization would be specifically useful for patients with chronic disease affecting several organs.

One of the key interests of such a tool is the possibility to build corpora of articles related to a given medical specialty.

5.4. Improving the results

The results clearly show that the performance of the CISMeF categorization algorithm is directly affected by existing gaps in our knowledge modeling. The lack of semantics relationships within the UMLS was a barrier in improving performances for Bodenreider [2]. However, we have the ability to enhance the coverage of our terminology. Based on the evaluation results, we have added 18 metaterms (and related semantic links) to the CISMeF terminology in March 2003, so that it now contains 85 metaterms.

We are currently planning further enhancement of the terminology through a second evaluation of the algorithm on another sample of 100 different resources. This new evaluation may enable us to find out whether it is necessary to create new semantic links or metaterms. It will also provide us with data to evaluate the impact of the recent additions to the terminology, and it will show whether these additions had significant influence on the *noise*.

Broader perspectives involve comparing the CISMeF categorization algorithm to more analytical categorization methods such as SVM and PPM compression based methods.

6. Conclusion

We have presented a categorization algorithm designed by the CISMeF team in order to make the resource description in the catalogue complete with a synoptic categorization of the medical specialties addressed in CISMeF in the form of a ranked list of relevant specialties. The automatic categorization method introduced in this paper is based on the manual indexing of resources with MeSH keywords/qualifiers pair and resource types. It also uses the semantic relationships between the different terms of the CISMeF terminology. The evaluation performed on 123 randomly picked resources gave very satisfying results, which enabled us to enhance the CISMeF terminology. The CISMeF team has decided that it was quite relevant to use this technique to generate resource categorization in the entire catalogue.

References

[1] T. Koch, Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways, Special issue of ''Online Information Review'' 24 (1) (2000) 24—34.

[2] O. Bodenreider, Using UMLS semantics for classification purposes, Proc. AMIA Symp. (2000) 86—90.

[3] S.J. Darmoni, J.P. Leroy, B. Thirion, F. Baudic, M. Douyère, J. Piot, CISMeF: a structured Health resource guide, Methods Inf. Med. 39 (1) (2000) 30—35.

[4] L.F. Soualmia, C. Barry-Greboval, H. Abdulrab, S.J. Darmoni, Modélisation et représentation des connaissances dans un catalogue de santé, Journées Francophones d'Ingénierie des Connaissances, IC'2002, Rouen, France, 2002, pp. 139—149.

[5] S.J. Nelson, D. Johnston, B.L. Humphreys, Relationships in medical subject headings, in: C.A. Bean, R. Green (Eds.), Relationships in the Organization of Knowledge, Kluwer Academic Publishers, New York, 2001, pp. 171—184.

[6] B. Thirion, S.J. Darmoni, Simplified access to MeSH tree structures on CISMeF, Bull. Med. Libr. Assoc. 87 (4) (1999) 480—481.

[7] P. Zweigenbaum, S.J. Darmoni, N. Grabar. The contribution of morphological knowledge to French MeSH mapping for information retrieval, Proc. AMIA Symp. (2001) 796—800.

[8] W. Kim, A.R. Aronson, W.J. Wilbur, Automatic MeSH term assignment and quality assessment, Proc. AMIA Symp. (2001) 319—323.

[9] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, J.G. Mork, S.J. Nelson, T.C. Rindflesch, W.J. Wilbur, The NLM Indexing Initiative, Proc. AMIA Symp. (2000) 17—21.

[10] A.R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, the NLM indexing initiative, Proc. AMIA Symp. (2001) 17—21.

[11] A. Gaudinat, C. Boyer, Automatic extraction of MeSH terms from Medline abstracts, in: NLPBA 2002—Workshop on Natural Language Processing in Biomedical Applications, MIE Special Topic Spring Conference, Nicosia, Cyprus, 8—9 March 2002, pp. 63—68, URL: http://www.genisis.ch/~natlang/NLPBA02/arnaud.pdf.

[12] S.J. Darmoni, E. Jarrousse, P. Zweigenbaum, P. Le Beux, F. Namer, R. Baud, J.M. Joubert, H. Vallée, R.A. Côté, A. Buemi, D. Bourigault, G. Recourcé, S. Jeanneau, J.M. Rodrigues, VUMeF: extending the French involvement in the UMLS metathesaurus, Proc. AMIA Symp. (2003) 824.

[13] W.J. Teahan, D.J. Harper, Using compression based language models for text categorization, in: J. Callan, B. Croft, J. Lafferty (Eds.), Workshop on Language Modelling and Information Retrieval, 2001, pp. 83—88.

[14] S.T. Dumais, Using SVMs for text categorization, in: M. Hearst (Ed.), IEEE Intelligent Systems Magazine Trends and Controversies, vol. 13 (4), 1998.

[15] L.M. Bruijn, A. Hasman, J.W. Arends, Automatic SNOMED classification—a corpus-based method, Comput. Methods Programs Biomed. 54 (1—2) (1997) 115—122.

[16] M. Mehta, R. Agrawal, J. Rissanen, SLIQ: a fast scalable classifier for data-mining, in: Proceedings of the 1996 International Conference on Extending Database Technologies (EDBT 496) Avignon, France, March 1996.

[17] J. Han, M. Kamber, Mining Text Databases, in: Morgan Kaufman Series (Eds.), Data Mining—Concepts and Techniques Data Management Systems, 2001, pp. 428—444.

[18] A. Wilcox, G. Hripcsak, Classification Algorithms Applied to Narrative Reports, Proc. AMIA Symp. (1999) 455—459.

[19] L.S. Larkey, W.B. Croft, Combining Classifiers in Text Categorization, in: Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, 1996, pp. 289—297.