eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

Mining Knowledge from Corpora: an Application to Retrieval and Indexing

Lina F.SOUALMIA[a,1], Badisse DAHAMNA[b], Stéfan DARMONI[b]
[a] *LIM&Bio EA 3969, SMBH Léonard de Vinci, Paris XIII University, Bobigny, France.*
[b] *CISMeF team and LITIS EA 4051,University of Rouen, France*

Abstract. The present work aims at discovering new associations between medical concepts to be exploited as input in retrieval and indexing. *Material and Methods*: Association rules method is applied to documents. The process is carried out on three major document categories referring to e-health information consumers: health professionals, students and lay people. Association rules evaluation is founded on statistical measures combined with domain knowledge. *Results*: Association rules represent existing relations between medical concepts (60.62%) and new knowledge (54.21%). Based on observations, 463 expert rules are defined by medical librarians for retrieval and indexing. *Conclusions*: Association rules bear out existing relations, produce new knowledge and support users and indexers in document retrieval and indexing.

Keywords. Data Analysis; Indexing; Terminology; Data Mining; MeSH.

Introduction

Internet is a major source of biomedical knowledge. As the access to structured medical information is difficult with directories or general search engines, many applications have been developed [1]. Since 1995, CISMeF (acronym of Catalog and Index of French-speaking Medical Sites) [2] has been selecting institutional and educational resources for patients, students and health professionals. It references 36,247 e-documents by using Medical Subject Headings (MeSH) [3]. Among many sources to support users, such as morphological bases, dynamic and contextual search tools [4], MeSH structure is exploited. To complete these sources, we propose to mine e-documents to discover new associations between medical concepts by data mining.

1. Material and Methods

1.1. Medical Subject Headings, Resource Types and Metaterms

The MeSH thesaurus is used by the National Library of Medicine for indexing biomedical resources. Its core is a hierarchical structure that consists of sets of

[1] Corresponding Author: Lina Soualmia, LIM&Bio EA 3969, 74 rue Marcel Cachin, 93017 Bobigny, France; E-mail: lina.soualmia@gmail.com.

descriptors: at the top level general headings (e.g. *diseases*) and deeper more specific headings (e.g. *brain infraction*). The 2007 version contains over 24,357 main headings (e.g.: *hepatitis*) and 83 subheadings (e.g.: *diagnosis*). Together with a main heading, a subheading can be used to specify a particular aspect. For example, the pair [*hepatitis/diagnosis*] specifies *diagnosis* aspect of *hepatitis*.

MeSH is originally used to index biomedical scientific articles for the MEDLINE database. In order to customize it to the field of e-health resources resource types have been introduced [2]. CISMeF resource types are an extension of MEDLINE publication types (e.g. *clinical guidelines*). Each document in CISMeF is described with a set of MeSH main headings, subheadings and CISMeF resource types. Each main heading, [main heading/subheading] pair and resource type is allotted a 'minor' or 'major' weight, according to the importance of the concept it refers to in the resource. Major terms are marked by a star (*).

1.2. Data Mining

Knowledge extraction from databases or data mining in computer science [5] consists in discovering additional information from large structured sets of data. This knowledge could be used to do predictions about new data or to explain existing data. One of the objectives of extraction process is the generation of association rules. It is processed in several steps: data and context preparation (objects and items selection), extraction of *frequent itemsets* (compared to a minimum support threshold), generation of *most informative rules* using a data mining algorithm, and finally interpretation and deduction of new knowledge [6]. An extraction context is a triplet $C=(O, I, R)$ where: O is the set of objects, I is the set of all the items and R is a binary relation between O and I.

1.2.1. Association Rules

A data mining system may generate several thousands and even several millions frequent association rules, and only some of them are interesting. An association rule is interesting if it is easily understandable by the users, valid for new data, useful or if it confirms a hypothesis. It is expressed as: $i_1 \wedge i_2 \wedge \ldots \wedge i_k \Rightarrow i_{k+1} \wedge \ldots \wedge i_n$ and states that if an object has the items $\{i_1,i_2\ldots,i_k\}$ it tends also to have the items $\{i_{k+1},\ldots,i_n\}$. *Support* represents the rule utility. It corresponds to the proportion of objects which contains at the same time antecedent and consequent. *Support* = $|\{i_1, i_2,\ldots, i_n\}|$. *Confidence* represents precision and corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent. Two rule types are distinguished: exact rule having Confidence=100%, i.e. verified in all the objects of the database and approximative rule. *Confidence* = $|\{i_1, i_2,\ldots, i_n\}|/|\{i_1, i_2,\ldots, i_k\}|$.

1.2.2. A-Close for Mining e-Documents

The problem of the relevance and the usefulness of extracted association rules is of a primary importance because real-life databases lead to several thousands and even millions of association rules whose confidence measures are high, and among which are many redundancies, i.e. rules conveying the same information among them. Two bases for association rules are defined by A-Close [7]. These bases generate sets for all valid non-redundant association rules, being thus smaller, composed by minimal antecedents and maximal consequents i.e. the most relevant association rules. We adapt A-Close to the case of e-health documents data base by considering conceptual indexing: the set of objects *O* is the set of indexed documents; the set of items *I* is the set of MeSH descriptors; the relation *R* represents the indexing relation between an object and an item, i.e. between a document and a descriptor.

1.2.3. Processing Collections of Documents

End-users are categorised in CISMeF in mainly three types: professionals, students in medicine, patients and lay people. Rather than extracting knowledge referring to the main medical specialties as in [4], we consider the three major resource types *guidelines**, *education** and *patients** and two kinds of itemsets: the set of major main headings (MH*) and the set of major [main heading/subheading] pairs (MH/SH*).

Table 1. Description of the collections of documents.

Resources	*Documents*	*Items*	*Min*	*Max*	*Mean*
Guidelines*	2,727	MH*	1	64	5.21
		MH/SH*	1	70	6.12
Patients*	3,272	MH*	0	25	1.63
		MH/SH*	0	30	1.82
Education*	3,610	MH*	0	25	2.22
		MH/SH*	0	34	2.73

2. Results

2.1. Mining e-Documents

For all contexts, minimum support was fixed to minsup=20 and minimum confidence to minconf=70% for the approximative association rules (Table 2). We obtain association rules between major MH* (resp. MH/SH* pairs). For the major resource types *patients** and *education** all (100%) association rules are between two MHs* (resp. MH/SH* pairs) i.e one descriptor in the antecedent and one descriptor in the consequent. For *guidelines** 24% of the rules are between more than two descriptors. Characteristics of documents may explain these results: average descriptors from 1.63 to 2.22 for *patients** and *education** whereas 5.21 to 6.12 for *guidelines**.

Table 2. Number of rules, exact rules (ER), approximative rules (AR) and pairs.

	*Context: item=MH**				*Context: item= [MH/SH]**			
Resources	*Rules*	*ER Conf=1*	*AR Conf≥0.7*	*Pairs*	*Rules*	*ER Conf=1*	*AR Conf≥0.7*	*Pairs*
Guidelines*	50	12 (24%)	38 (76%)	38 (76%)	39	8 (20.51%)	31 (79.49%)	35 (76%)
Patients*	20	9 (45%)	11 (55%)	20 (100%)	19	8 (42.1%)	11 *(57.9%)*	19 (100%)
Education*	23	6 (26.09%)	17 (73.91%)	23 (100%)	25	13 (52%)	12 (48%)	25 (100%)

Another experiment is carried out in the context of documents with the resource type *guidelines** to obtain more complete association rules: we consider the descriptors MH and MH/SH pairs without alloted minor or major weight. An average of 12 descriptors with a minimum of 1 and a maximum of 301 descriptors compose the documents (Table 3). As A-Close works on databases with a maximum of 12 items, we have added a constraint on the number of descriptors. To avoid long time generation and to have interpretable association rules, we added the maximum size of the closed itemsets as a new parameter of the algorithm.

Table 3. Description of the documents of the Guidelines* collection.

	Docs	*Items*	*Min*	*Max*	*Mean*
Guidelines*	2,727	MH	1	111	10.08
		MH/SH	1	301	13.54

We obtain a high number of association rules with a minimum support threshold *minsup*=20 and a minimum confidence threshold *minconf*=70% (Table 4) but only 0.95% (respectively 1.92%) are between two MH (respectively between two MH/SH pairs). By reducing the confidence from 1 to 0.7 the number of rules between MH (respectively between MH/SH) growths with a factor of 5 (respectively 4.42).

Table 4. Association rules between MH and MH/SH in the context Guidelines*.

Items	*Rules*	*ER* *Conf=1*	*AR* *Conf≥0.7*	*Pairs*
MH	35,454	6,990 (19.71%)	28,464 (80.29%)	338 (0.95%)
MH/SH	27,011	6,102 (22.6%)	20,909 (77.4%)	520 (1.92%)

2.2. Association Rules Evaluation

As defined, an interesting association rule confirms a hypothesis or states a new hypothesis [6]. We propose here to combine background domain knowledge with simple statistical measures used traditionally in association rules mining for evaluation. We consider several cases of interesting association rules according to relations between MeSH descriptors [3]. As these relations are defined between two main headings and between two subheadings we consider only the association rules between two elements. Hence, an interesting existing association rule could associate: a (in)direct son and its father (FS); two descriptors that belong to the same hierarchy (same (in)direct father) (B); two descriptors with See Also relation (SA). These rules are automatically classified thanks to MeSH structure. The other rules that satisfy the *misup* and *minconf* are then considered as «new» interesting association rules.

Exact association rules, except for collection *patients**, are mostly new interesting rules: from 62.5% to 99.86%. Therefore, existing rules are mainly from the *patients** collection: 77.77% for MH* and 75% for MH/SH*. Approximative rules, except for the *guidelines** collection with items MH and MH/SH pairs, are mostly existing interesting rules: from 58.07% to 78.73%. New interesting rules are between MH and MH/SH from the collection *guidelines**: 99.73% for MH and 99.52% for MH/SH.

Subjective interest measures are based on the expert knowledge about the data, i.e. here the medical librarian. New interesting rules for the contexts MH* and MH/SH*

Figure 1. Some examples of new interesting rules validated by the expert

breast cancer/diagnosis → mammography
aids/prevention and control → condom
influenza vaccines→influenza/prevention and control
Turner syndrome ∧ child → human growth hormone ∧ growth disorders
obstetric delivery→ pregnancy
prostate cancer/surgery→ biopsy ∧ prostatectomy
amniocentesis → prenatal diagnosis ∧ chorionic villi sampling
opioids analgesics/administration and dosage → pain/drug therapy

pairs are evaluated manually. 93.75% (resp. 84.78%) of the interesting new rules with confidence=1 (resp. confidence≥0.7) between major descriptors are validated.

Table 5. Association rules evaluation according to MeSH structure.

		Exact rules: Confidence=1				***Approximative rules: Confidence≥0.7***			
		Existing knowledge			*New*	*Existing knowledge*			*New*
		FS	*B*	*SA*		*FS*	*B*	*SA*	
Patients*	MH*	0 0%	5 55.55%	2 22.22%	2 22.22%	2 18.18%	2 18.18%	4 36.36%	3 27.27%
	MH/SH*	0 0%	5 62.5%	1 12.5%	2 25%	2 18.18%	2 18.18%	3 27.27%	7 36.36%
Education*	MH*	1 16.66%	1 16.66%	0 0%	4 66.67%	2 11.76%	6 35.29%	3 17.64%	6 35.29%
	MH/SH*	1 7.69%	0 0%	1 7.69%	11 87.62%	2 16.76%	3 25%	2 16.76%	5 41.66%
Guidelines*	MH*	0 0%	0 0%	4 33.33%	8 66.67%	2 5.26%	7 18.42%	10 26.31%	12 31.57%
	MH/SH*	1 12.5%	1 12.5%	1 12.5%	5 62.5%	3 9.67%	3 9.67%	9 29.03%	13 41.93%
	MH	0 0%	2 0.03%	8 0.14%	6,980 99.86%	12 0.04%	37 0.13%	30 0.10%	28,382 99.73%
	MH/SH	6 0.1%	4 0.06%	7 0.11%	6,085 99.73%	25 0.12%	50 0.23%	27 0.13%	20,807 99.52%

2.3. Indexing Correction and Expert Rules

Documents are manually indexed and according to the indexing policy, the more precise descriptor should be used, i.e. in lower level in hierarchy. However, 1,466 documents contain descriptors that have father-son relation and 478 documents are indexed by subheadings that have a relation while associated to the same keyword. For example, a document is indexed by *trisomy* and *chromosome aberrations*, whereas *trisomy* is a *chromosome aberration*. This may explain the proportion of existing associations. Correction should be proposed to the indexers.

Main return on experiences of association rules extraction and evaluation is modeling and formalisation of rules between [main heading/subheading] pairs based on observations. The pattern of the rule *hepatitis/prevention and control→hepatitis vaccines* is used to model *dysentery bacillary/prevention and control → shigella vaccines*. 463 rules are modeled. Formalization concerns different cases and contexts for retrieval and indexing. The rule $MH_1/SH_1 \xrightarrow{--} MH_2$ states that MH_1/SH_1 should be replaced by the main heading MH_2. For example *abdomen/radiography* $\xrightarrow{--}$ *radiography abdominal*. The rule $MH_1/SH_1 \xrightarrow{++} MH_2/SH_2$ states that the pair MH_2/SH_2 should be added to the pair MH_1/SH_1. *appendectomy* $\xrightarrow{++}$ *appendicitis/surgery* states that the pair *appendicitis/surgery* should be added to queries (or to document description when indexing) containing the main heading *appendectomy*.

3. Discussion and Future Work

There is an increasing activity in text mining in the genomic model [8]. In [9] co-occurrences between Gene Ontology terms are analyzed and association rules are mined to identify pairs of related Go Terms. Association rules are more complete than co-occurrences measures between pairs of concepts but one of the challenging issues is the overabundance of associations that may be discovered as in [10]. A-Close generates all the valid non-redundant association rules composed by minimal antecedents and maximal consequents. Evaluation is processed in two steps: first the selection of the most informative rules and second the classification of the rules according to the MeSH taxonomy structure to filter existing associations. Only the most frequent rules that are not classified are presented to the expert for a final evaluation. This method combines statistical measures and background domain knowledge.

Association rules are used in retrieval by query expansion (automatic and interactive) and enriching users' queries with new knowledge [4]. As exact rules (respectively approximative rules) state that the antecedent and the consequent are at the same time in all (respectively some) documents, this kind of rules should be used in automatic (respectively interactive) query expansion. However, these expansions work only in the case of queries that return documents. Association rules link conceptual structures of the documents i.e. descriptors organised in hierarchies on which it is possible to make specialization and generalization. We plan to generate generalized association rules and to examine how other data collections such as MEDLINE will work with our approach. Association rules and expert rules can be translated in the form of automatas for processing automatic indexing of raw text documents. Finally formalised association rules could improve the power of reasoning based on MeSH-OWL [11].

References

[1] Abad Garcia F et al. A comparative study of six European databases of medically oriented Web resources. *J Med Libr Assoc.* 2005;93(4):467-479.

[2] Douyère M et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J.* 2004;21(4):253-261.

[3] Nelson SJ et al. Relationships in MeSH. *Kluwer Publishers* 2001;171-84.

[4] Soualmia LF, Darmoni SJ. Combining knowledge-based methods to refine and expand queries in medicine. *LN in Computer Science* 2004;3055:243-255.

[5] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. *Very Large Data Bases* 1994;478-499.

[6] Fayyad UM et al. Advances in knowledge discovery and data mining. Am. Ass. Artificial Intelligence Press 1996;601-611.

[7] Lakhal L et al. Efficient mining of association rules using closed itemset lattices. *Information Systems* 1999;24:25-46.

[8] Ananiadou S, Mc Naught J. Text mining for biology and biomedicine. *Artech House publishers* 2005.

[9] Bodenreider O et al. Non lexical approaches to identifying associative relations in the Gene ontology. *Pacific Symposium on Biocomputing* 2005;10:91:102

[10] Berardi M et al. A data mining approach to PubMed query refinement. DEXA 2004. *IEEE Computer Society*; 401-405.

[11] Soualmia LF et al. Representing the MeSH in OWL: towards a semi-automatic migration. KR-Med 2004;72-80.