BNDI:A Baysien Network for biomedial Documents Indexing with MeSH thesaurus

Wiem CHEBIL[1,2], Lina F SOUALMIA[2], Mohamed N OMRI[1] and Stephan J DARMONI[2]

[1]MARS Research Unit,
Faculty of sciences of Monastir,
Avenue of the environment, 5019 Monastir, Tunisia

[2]Normandie Univ., CISMeF Team,
LITIS-TIBS EA 4108, Rouen University and Hospital
Wiem.chebil@yahoo.fr, Lina.Soualmia@chu-rouen.fr
Mohamednazih.omri@fsm.rnu.tn, Stefan.Darmoni@chu-rouen.fr

***Abstract*—We propose a new approach denoted BNDI (Bayesian Network for Document Indexing) for indexing biomedical documents with controlled biomedical vocabulary based on a Bayesian Network. BNDI uses the probability inference to extract descriptors from the document. The main contribution of our approach is that it takes into account the structure of biomedical terminologies specially The Medical Subject Heading (MeSH) thesaurus. We experimented our approach on a subset of the OSHUMED collection.**

Key words: Bayesien Network; Biomedical documents; terms extraction, MeSH thesaurus;

1. INTRODUCTION:

The number of biomedical resources in the Internet is in permanent increase which makes the task of the human indexer more difficult. To lead with this issue, several approaches of biomedical document indexing were proposed. In this paper, we proposed a new unsupervised Bayesian network-based approach for indexing biomedical document with a biomedical controlled vocabulary. Dealing with the fact that the task of assigning relevant descriptors to the document is uncertain, we choose to use probabilistic Bayesian Networks (BN). A BN is a robust inference mechanism for reasoning under condition of uncertainty. To the best of our knowledge, there is only one Bayesian Network-based approach that have been proposed by De compos *et al.* [1] for indexing document that showed good results that out-performed the Vector Space Model (VSM) method (the Average 11-point precision of VSM is 0.17 vs. the Average 11-point precision of De Campos *et al.* is 0.34). However the approach of De Campos *et al.* doesn't use the information provided by the frequency of descriptors words in the document which is pertinent information that contributes to estimate the relevance of a descriptor given a document. In addition, in [1] the structure of a thesaurus that has been considered is not adequate to the structure of medical thesaurus. For example the authors of [1] consider that there are only equivalent relations between the entry term and the descriptors. These relations are "genuine synonymy, near-synonymy, antonymy and inclusion, » which is not the case for the Medical Subject Headings Thesaurus (MeSH) [2]. In fact, only the synonymy relation among these relations and other hierarchical relations linked the entry terms and the descriptors in MeSH (more details about the structure of MeSH are in section 3). In the other hand, the approach proposed in [1] favors the generic descriptor not the specific ones. However the more a descriptor is specific the more it is precise for indexing. So our main contribution is to deal with these limitations. This paper is organized as follow: the section 2 presents the related

work, in the section 3 we define the MeSH thesaurus and the BN. We detail, in the section 4 the proposed approach. The experimentations are explained in section 5. The section 6 presents the results and the discussion. Finally, in section 7 we conclude and we cite our future work.

2. RELATED WORK:

Pouliquen *et al.* [3] computed a statistic weight based on Term Frequency-Inverse Document Frequency TF-IDF for each term automatically extracted from the document using a method based on Natural Language Processing (NLP). These terms are then matched to the terms of the ADM (assistance with the medical diagnosis) dictionary. Jonquet *et al.* [4] applied the Mgrep tool for extracting concepts from 200 biomedical ontologies, and computed a score for each generated annotation according to its origin (preferred term, non-preferred term, synonym term …etc.).. Mukherjea *et al.* [5] developed BioAnnotator a tool for indexing biomedical documents. It uses a parser to identify noun phrases from a document and then matches them to the UMLS concepts using a rule engine. Zhou *et al.* [6] proposed to annotate documents with only the most significant words in the UMLS Meta-thesaurus. Ruch [7] proposed an indexing approach denoted by Eagl that combined two models: the Vector Space Model (VSM) and a regular expression pattern matcher. We can also cite the work of Majdoubi *et al.* [8] that used VSM to extract MeSH terms and then computed a statistic and semantic weight for ranking these terms. The indexing technique of Aronson *et al.* [9] is based on three methods. The first matches the document terms with UMLS terms using MetaMap (software tool for English that allows mapping document to the concepts UMLS). The second method compares the document phrase and the concept phrase using tri-gram method. The third method extracts MeSH terms from the k-nearest neighbors of the indexing documents and ranks them using a statistic weight. Couto *et al*. [10] computed likelihood between Gene Ontology terms and a document using the Evidence Content (EC) of a term, which is the sum of all EC of its words. The EC of a word is its weight in the ontology. Duy *et a.l* [11] combined VSM with a proposed similarity between terms and the document that takes into a count the words order in the document. Hliaoutakis *et al*. [12] proposed Automatic MeSH Term Extraction (AMTx) model that uses C/NC-value method that allows the extraction of the multi-word terms from a document by combining statistic and linguistic information. The candidate terms are those that correspond to MeSH terms.

3. BACKGROUND

A *The MeSH thesaurus*

The MeSH thesaurus® [2] is a controlled vocabulary created by the US. NLM and it is used for indexing medical resources such as MEDLINE which is one of the biggest bibliographic medical databases. The MeSH is composed of: main headings, subheadings and supplementary chemistry concept (CCSs). The main headings (or descriptors) are used for description of the medical articles and indexing citations. Each descriptor consists of entry terms[1] . There are 26,142 descriptors and over 177,000 terms in 2011. The sub headings define the meaning of a descriptor. The CCCs index the chemistry product and the drugs.

B *The Bayesian Network*

Bayesian Networks (BN) (or belief networks) are graphical tools that allow reasoning in a specific under uncertainty quantitatively and qualitatively in a specific domain and allow also to make inferences. A graph of a BN G (V,L) is directed and acyclic, composed of nodes which represent random variables V(A,B,..) and linked with a causal dependency L=V ×V over which is defined a probability distribution.

- A link from variable A to variable B indicates that A can cause or affect B. A is a "parent of" B, and B is a "child of" A (ancestor descendant).
- A nod without parent is called a root node.
- A node without children is called a leaf node.
- An other node (non-leaf and non-root) is called a intermediate node
- A chain is a sequence of nodes linked with arcs.

4. DESCRIPTION OF THE PROPOSED APPROACH:

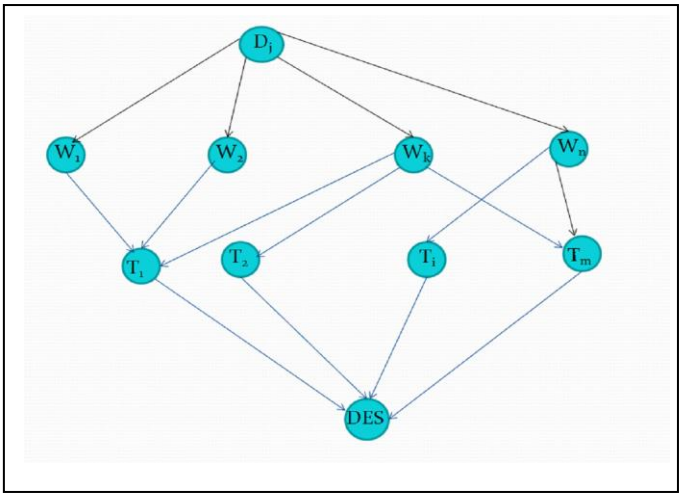
Our approach for indexing biomedical documents is composed of two steps : the first step is the pretreatment of the document as well as the MeSH terms and the second step is the extraction of descriptors using a BN.

A Pretreatment:

The step of pretreatment consists of 3 tasks: (i) removing punctuation (ii) pruning stop words (iii) stemming. These tasks are applied on document and MeSH terms. Let "The binding of acetaldehyde to the active site of ribonuclease: alterations in catalytic activity and effects of phosphate." a title of a document. After the pretreatment this title become "bind acetaldehyd activ site ribonucleas alter catalyt activ effect phosphat". We chose to use PORTER Algorithm [13] for stemming. PORTER is based on rules that allow removing words suffix and reduce a word to its stem (or root).

B Extraction of descriptors using Bayesian Network:

Figure1: The Bayesian network for indexing biomedical documents



1) The graphical components

The graphical component (see figure 1) represents the document D_j, the word of the text M_k, the term MeSH T_i and the descriptor DES nodes and the (in) dependence relations existing between nodes. If the document is instantiated it means that *Dj = dj* . *Dj = ¬dj* if the document is not instantiated. We are only interested to the case when the document is instantiated and we noted indifferently Dj. Wi references a word in the text or in a term. The domain of words is dom(W_k) = {w_k, $\neg w_k$}. If the word occurs in the text $W_i=w_k$, if the word is absent in the text $W_k=\neg w_k$. The domain of the terms is dom(T_i) = {t_i, $\neg t_i$}. The value of T_i is $T_i = t_i$ if t_i is a term of the instantiated descriptor. $T_i = \neg t_i$ if not. We consider only the terms of the instantiated descriptors. The domain of a descriptor is dom(DES) = {des, ¬des}. Des=des if the descriptor is instantiated DES= ¬des if not. We interest only to the case where the descriptor is instantiated and we denoted DES indifferently.

2) Evaluation of a descriptor:

The aim of our proposed model is to compute the probability of the relevance of a descriptor given a document. When a document from the collection is instantiated, the distribution of the information is activated by this instantiation from the document to the descriptor. For each node the a posteriori conditional and marginal probability is computed being given the a priori conditional and marginal probability.The conditional probability of a node depends on all the possible configurations of its parents.

a) Aggregation of terms words:

Turtle in his Bayesian Network Information Retrieval (IR) model proposed five canonical forms for each type of search [14]. These forms can be adopted in our model by replacing the query by the term. Thus, a term can be aggregated by the Boolean operators (OR, AND, NOT) and probabilistic sum or one of its variations the weighted sum. To evaluate the conditional probabilities P(T| θ) (θ is all the set of parents of T) of a node T having n parents {θ_1, .., θ_n}, and $P(\theta_1 = w_1) = p_1$, .., $P(\theta_n = w_2) = p_n$ the following aggregations are defined (1):

$$
\begin{aligned}
&P_{or}(T \mid \theta) = 1 - (1 - p_1) - .. - (1 - p_n)\\
&P_{and}(T \mid \theta) = p_1 \times .. \times p_n\\
&P_{Not}(T \mid \theta_1) = 1 - p_1\\
&P_{sum}(T \mid \theta) = p_1 + .. + p_n\\
&P_{weighted\ sum}(T \mid \theta) = (wt_k\, p_1 + wt_k\, p_i .. + wt_n\, p_n)\, wt_t
\end{aligned}
\qquad (1)
$$

Wt_k is the weight of the word w_k
W_t is the weight of the term

In our approach, we consider that all terms words must occur in the text, thus the term is aggregated by the Boolean operator AND (a conjunctive aggregation).

b) The conditional probability of a descriptor given his terms:

To compute the conditional probabilities of a descriptor node *D* given his terms we based on the assumption that the more a descriptor have terms in the document and these terms are frequent the more it is relevant. For that, we used the canonical additive model proposed by De Campos et al. [1]. Thus, the conditional probability of a descriptor given his terms is equal to the sum of the weight of the instantiated terms. This weight is equal to the weight of the relation that linked the term and the descriptor. This sum is divided by the number of the instantiated terms in the document to ensure that the probability of the descriptor given his terms is not upper than one.

c) Weighting the arcs $P(w_k \mid D_j)$

The arc that links the words to the document is weighted using wf idf.. (word frequency inverse document frequency)
Thus :

$$P(w_k \mid D_j) = wf_{ij} * idf_i \quad (2)$$

$$wf_{ij} = \frac{freq_{ij}}{\max_{r:1->n}(freq_{rj})} \quad (3)$$

$$idf = \log\left(\frac{n_i}{N}\right) \quad (4)$$

N is the number of documents in the collection and n_i is the number of documents where w_i occurs.
n: is the number of word in the document
$freq_{i\,j}$: the frequency of a word i in a document j

d) The probability of the descriptor given a document:

According to the graph and using the aggregation of terms words , the conditional probability of a descriptor given his terms as well as the weight of arcs $P(M_k \mid Dj)$ which are defined in the previous sections, we define the probability of a relevance of a descriptor given a document as follow :

$$P(DES \mid D_j) = p(DES \mid \theta^t)\, p(\theta^t \mid D_j) \quad (5)$$

$$p(des \mid \theta^t) = \frac{\sum_{t_i \in \theta^t} P(DES \mid t_i)}{m} \quad (6)$$

$$\text{With } P(DES \mid t_i) = W(DES \mid t_i) \quad (7)$$

$$P(\theta^t|DES) = \prod_{t_i \in \theta^t} \left(\prod_{W_k \in W(T) \wedge W(d_j)} \left(P\left(w_k \middle| D_j \right) \right) \right) (8)$$

θ^T : all the possible configurations of the set of parents of DES.

θ^t : is a possible configuration in θ^T .

The possible configuration of the descriptor terms (the parents of DES) composed of the terms {T1, T2} are

$\theta^T = \theta^t$ = {{t_1, t_2}} (as explained in the section 1))

W(DES |t_i) : is the weight of the relation between descriptors and terms. We suppose that all the relations have the same weight which is equal to 1. If t_i have at least one parent that doesn't occur in the document the W(DES |t_i) is equal to 0.

m : is the number of a descriptors terms

5. EXPERIMENTATIONS

To test our approach we selected 6,000 citations among the OHSUMED collection composed of 4,591,015 MEDLINE citations. Each selected citation is composed of title and an abstract. The content of the title is merged with the content of the abstract when indexing the citations. We compared the automatic index with the manual one which is considered as the gold standard. We computed the precision (P), recall (R) and Fscore (Fs). The precision is the Number of Correct Descriptors Automatically Extracted (NCDAE), devised by the Total Number of Descriptors Automatically Extracted (TNDAE). The recall is the number of correct descriptor devised by the Number of Descriptors Manually Extracted. F-score combines precision and recall with an equal weight [22]. In order to compare our approach to other one we evaluated in the step of experiments MTI [10] and Eagle [12] which were described in section 2.

$$\text{Precision} = \frac{\text{NCDAE}}{\text{TNDME}} (9)$$

$$\text{Recall} = \frac{\text{NCDAE}}{\text{NDME}} (10)$$

$$F - score = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precision} + \text{Recall}} (11)$$

For example, let the MeSH descriptors: Acetaldehyde, Microfilament Proteins, and the following document having the PMID= 2789522 (see figure 2):

Title: Covalent interactions of acetaldehyde with the action/microfilament system

Abstract: The covalent binding of [14C] acetaldehyde to purified rabbit skeletal muscle actin was characterized. As we have found for other cytoskeletal proteins,actin formed stable covalent adducts under reductive and non-reductive conditions. Under non-reductive conditions, individual and competition binding studies versus albumin both showed that the G-form of actin is more reactive toward acetaldehyde than the F-form. When proteins were compared on an 'equi-lysine' basis under non-reducing conditions, G-actin was found to preferentially compete with albumin for binding to acetaldehyde. Time-course dialysis studies indicated that acetaldehyde-actin adducts become more stable with prolonged incubation at 37 degrees C. These data raise the possibility that actin could be a preferential target for adduct formation in cellular systems and will serve as the basis for ongoing studies aimed at defining the role of acetaldehyde-protein adducts in ethanol-induced cell injury.

Figure 2 : Example of MEDLINE citation

P(Acetaldehyde /d_j)=w(Acetaldehyde/ Acetaldehyde) ×p(Acetaldehyde / d_j)=1*1=1

P(Microfilament Proteins /d_j)=w(Microfilament Proteins/ Microfilament Proteins) ×p(Microfilament /d_j) p(Proteins/d_j)=1*(1/6)*(3/6)=0.08

Thus, the descriptor « Acetaldehyde" is ranked before the descriptor "Microfilament Proteins" for indexing the document.

6. RESULTS

Table 1: The recall of BNDI, Eagle and MTI

Rank	R-BNDI	R-Eagle	R-MTI
1	0.24	0.25	0.25
5	0.37	0.38	0.37
10	0.46	0.48	0.47
15	0.52	0.55	0.52

Table 2 : The precision of BNDI, Eagle and MTI

Rank	P-BNDI	P-Eagle	P-MTI
1	0.72	0.61	0.70
5	0.60	0.45	0.56
10	0.49	0.30	0.45
15	0.41	0.25	0.38

Table3 : The F score of BNDI, Eagle and MTI

Rank	Fs-BNDI	Fs-Eagle	Fs-MTI
1	0.36	0.35	0.36
5	0.45	0.41	0.44
10	0.47	0.36	0.45
15	0.45	0.34	0.43

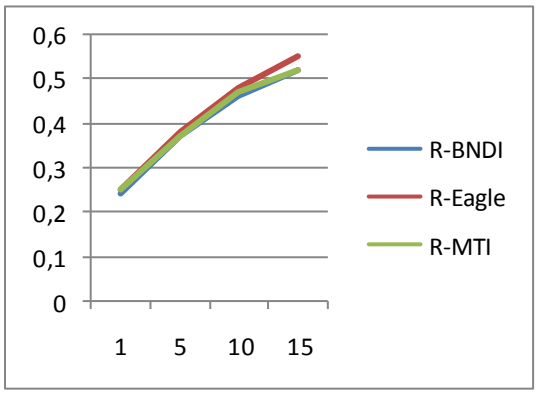


Figure 3: Variation of the recall according to the rank.

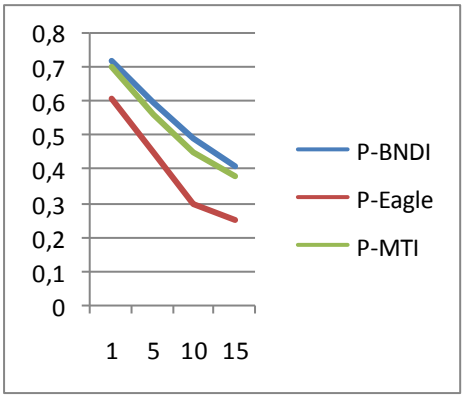


Figure 4: Variation of the precision according to the rank.

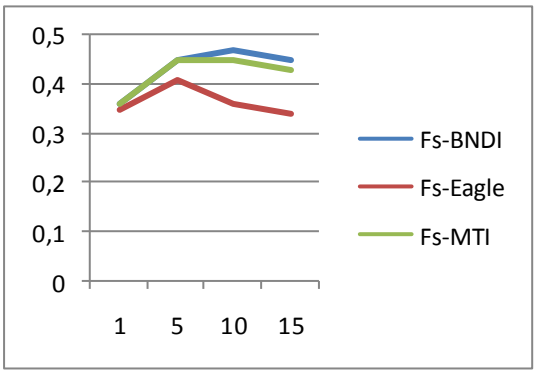


Figure 5: Variation of the f-score according to the rank.

Table 1 and figure3 show that in term of recall and at rank 5 and 15 our approach and MTI have the same recall while Eagle outperforms the both (MTI and our approach) at 15. In addition, MTI and Eagle outperform our approach at ranks 1 and 10. According to the table 2, and the figure 4 we can note that the precision of our method is upper than the precision of MTI and Eagle at different ranks. We can explain these results by the fact that our approach allows extracting terms having all their words occurring in the document while Eagle allows extracting terms having at least one word don't occur in the document and MTI lead to extract terms from citations similar to the citation being indexed. Moreover, when we analyse the table 3 and the figure 5 we can see that f-score of our approach is also better then the other approaches. Thus, we can conclude according the results specially f-score and comparing to two other approaches that the performance of our proposed approach is quite acceptable.

7 .Conclusion and future work:

We presented in this paper a new approach for indexing biomedical documents with MeSH thesaurus based on a Bayesian Network. Our main contribution comparing to [1] that our approach takes into account the occurrences of terms words in the text which is an important element that contribute to estimate the relevance of terms given a document. Moreover, the Bayesian network that we developed takes into account the structure of biomedical terminologies, the MeSH specifically. The experimentations of our approach on the OSHUMED corpus show that the Bayesian Network model is adequate for indexing biomedical documents. For this reason we aim to apply this model on more than one terminology and on other corpora such as CISMeF corpus[2].

References

[1] De Campos LM, Fernández-Luna JM, Huete JF, and Romero AE. *Automatic Indexing from a Thesaurus Using Bayesian Networks: Application to the Classification of Parliamentary Initiatives.* 9th European Conference, ECSQARU 2007, Hammamet, Tunisia, 2007.

[2] Nelson SJ, Johnson WD, Humphreys BL. *Relationships in Medical Subject Heading. In: Relationships in the Organization of Knowledge*, 2001, eds. Kluwer Academic Publishers, pp. 171–184(2001)

[3] Happe, A, Pouliquen B, Burgun A, Cuggia M, Beux P.L. *Automatic concept extraction from spoken medical reports*. I.J medical Informatics; 70(2-3): pp255-63(2003).

[4] Jonquet C, LePendu P, Falconer SM, Coulet A, Noy N.F, Musen M.A, Shah N.H. *NCBO Resource Index: Ontology-based search and mining of biomedical resources*. J. web Sem; 9(3): pp316-324 (2011).

[5] Mukherjea et al. *Enhancing a biomedical information extraction system with dictionary mining and context Disambiguation*. IBM Journal of Research and Development; 48(5/6): pp 693-701(2004).

[6] Zhou X, Zhang X, Hu X. MaxMatcher: *biological concept extraction using approximate dictionary lookup*. PRICAI Pacific Rim International Conferences on Artificial Intelligence; pp145–149(2006).

[7] Ruch P. *Automatic assignment of biomedical categories: toward a generic approach*. Bioinform J; 22(6):658–64(2006).

[8] Majdoubi J, Tmar M, Gargouri F. *Using the MeSH thesaurus to index a medical article: combination of content, structure and semantics.* International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES (1):277–84(2009).

[9] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. *The NLM indexing initiative's medical text indexer.* Med Health Info; 11(1):268–72 (2004).

[10] Couto.F M., Silva M.J, Coutinho. *Finding genomic ontology terms in text using evidence content*. BMC Bioinformatics 2005; 6:(S-1).

[11] Dinh, D., Tamine, L.: *Biomedical concept extraction based on combining the content-based and word order similarities*. Symposium On Applied Computing SAC; 1159-63(2011).

[12] Hliaoutakis A, Zervanou K, Petrakis EGM. *The AMTEx approach in the medical document indexing and retrieval application*. Data Knowledge Eng. 2009;68(3):380–9

[13] Porter,M. *An algorithm for suffix stripping*, Program 1981; 14(3):130-137.

[14] Turtle, H. *Inference networks for document retrieval*, 1991. Ph.D thesis university of Masschutts

[2] http://www.chu-rouen.fr/cismef/