JHI JOURNAL OF HEALTH INFORMATICS

Similarity-based scoring method for classification of Health Informatics content

Método baseado no escore de similaridade para a classificação de conteúdo em Informática em Saúde

Método basado en la puntuación de similitud para clasificar el contenido en Informática de la Salud

Fabio Teixeira[1], Alex Jaccoud Falcão[1], Fernando Sequeira Sousa[1], Anderson Diniz Hummel[1], Thiago Martini Costa[1], Felipe Mancini[2], Luciano Vieira de Araujo[3], Ivan Torres Pisa[1]

Keywords: Classification; Artificial Intelligence; Data Analysis

ABSTRACT

Objective: There has been a considerable growth of the architecture and complexity of digital repositories in Health Informatics (HI). For information retrieval different information treatment and representation, such as automatic content classification, are required. The purpose of this study is to present the results of a procedure for automatic classification of scientific articles in HI using a specific thesaurus. **Design:** Statistical, vector, and artificial intelligence methods were applied to classify HI-related content. Articles extracted from the HI and Health journals and a specialized HI thesaurus were used for method application and result evaluation. **Measurements:** Statistical procedures and measures of accuracy, precision, recall, area under the ROC curve, and combination of precision and recall (F_1 measure) were performed to measure the degree of similarity between terms of the specialized HI thesaurus and the selected articles. **Results:** The percentage of accuracy achieved was 0.87, F_1 measure was 0.87 and the area under the ROC curve was 0.94. **Conclusion:** The results were positive, showing that the use of a specialized thesaurus on Health Informatics in conjunction with the methods used allows the classification of articles in the areas of Health Informatics and Health.

Descritores: Classificação; Inteligência artificial; Análise de dados

RESUMO

Objetivo: Há um crescimento considerável na arquitetura e complexidade dos repositórios digitais em Informática em Saúde (IS). A recuperação de informação neste cenário requer diferentes tratamentos e representações, como a classificação automática de conteúdo. O propósito deste estudo é apresentar os resultados de um processo automatizado para a classificação de artigos científicos de Informática em Saúde, utilizando um tesauro especializado neste domínio de conhecimento. **Métodos:** Métodos estatísticos, vetoriais e de inteligência artificial foram aplicados para classificar conteúdo relacionado à Informática em Saúde. Artigos científicos publicados em revistas de Saúde e Informática em Saúde, bem como um tesauro especializado em Informática em Saúde foram utilizados para a aplicação dos métodos e avaliação dos resultados.**Avaliação:** Métodos estatísticos e medidas de acurácia, precisão, revocação, área sob a curva ROC e F_1-measure foram realizadas para medir o grau de similaridade entre os termos do tesauro especializado e os artigos selecionados. **Resultados:** O percentual de acurácia obtido foi de 0.87, F_1-measure foi 0.87 e a área sob a curva ROC foi 0.94. **Conclusão**: Os resultados obtidos foram positivos, mostrando que a utilização de um tesauro especializado em Informática em Saúde em conjunto com os métodos aplicados possibilita a classificação de artigos nos domínios da Informática em Saúde e Saúde.

Descriptores: Clasificación; Inteligencia artificial; Analisis de datos

RESUMEN

Objetivo: Hay un aumento considerable de la complejidad y la arquitectura de los repositorios digitales en Informática de la Salud (IS). La recuperación de la información en este escenario requiere diferentes tratamientos y actuaciones, como la clasificación automática de contenidos. El propósito de este estudio es presentar los resultados de un proceso automatizado para la clasificación de artículos científicos sobre Informática en Salud, utilizando un diccionario de sinónimos en la misma área de interés. **Métodos:** Los métodos estadísticos, el vector y la inteligencia artificial han sido aplicados para clasificar los contenidos relacionados con la Informática en Salud. Artículos publicados en revistas de Salud y de Informática en Salud, así como un diccionario especializado en Informática en Salud se utilizó para la aplicación de métodos y la evaluación de los resultados. **Clasificación:** Métodos estadísticos y medidas de la exactitud, precisión, cobertura, área bajo la curva ROC y F1 mediciones se realizaron para medir el grado de similitud entre los términos del diccionario de sinónimos y artículos especializados seleccionados. **Resultados:** El porcentaje de precisión obtenido fue de 0,87, F1-medida fue de 0,87 y el área bajo la curva ROC fue de 0,94. **Conclusión:** Los resultados fueron positivos, demostrando que el uso de un tesauro especializado en Informática en Salud en relación con los métodos que permite la clasificación de los artículos en las áreas de Informática en Salud y Salud.

[1] Departamento de Informática em Saúde,Universidade Federal de São Paulo - UNIFESP - São Paulo (SP), Brasil.

[2] Departamento de Informática em Saúde,Universidade Federal de São Paulo - UNIFESP - São Paulo (SP), Brasil. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP - Guarulhos (SP), Brasil.

[3] Escola de Artes, Ciências e Humanidades, Universidade de São Paulo - USP - São Paulo (SP), Brasil.

Autor Correspondente: **Ivan Torre Pisa**
e-mail: **buscasaude@unifesp.br**

Artigo recebido: **03/02/2011**
Aprovado: **01/04/2011**

INTRODUCTION

The expansion of the World Wide Web has stimulated the growth of the architecture and complexity of digital repositories in health informatics (HI). For information retrieval different information treatment and representation, such as automatic content classification using controlled vocabularies, are required[1].

The development of several research areas, including HI, Information Sciences, Artificial Intelligence, Information Retrieval, and Computational Linguistics has promoted the study and use of terminologies such as controlled vocabularies, subject headings, and thesauri[2].

Terminologies are used as an instrument for information indexing, classification, and retrieval, and the thesaurus is the most sophisticated and commonly used tool[3].

A thesaurus is defined as a controlled vocabulary representing hierarchies, equivalence, belonging relations, and associations between terms designed to help users find the information they need with the narrowest margin of error[4].

In this context, a specialized HI thesaurus allows to better guide users in their search of specific contents relevant to its area of application, encouraging its use while building automated tools for content classification and information retrieval.

The objective of the present study is to classify HI-related scientific articles using statistical, vector, and artificial intelligence methods for fast and effective automated identification, finding, and retrieval of information based on a specialized HI thesaurus.

BACKGROUND

There is a large body of literature on text classification in scientific papers available at Medline (U.S. National Library of Medicine's premier bibliographic database). For example, Torvik[5] works on a model for estimating the probability that a pair of author names, sharing last name and first initial, appearing on two different Medline articles, refer to the same individual. Other initiatives, like Bolelli[6], investigate temporal relationships between scientific literature collections that can help the topic discovery process.

In addition, The US National Library of Medicine (NLM) supports the Text Categorization (TC) project, for encouraging researchers to develop tools that can automatically index health-related texts based on a controlled vocabulary[7].

Text Categorization Project

This research project is subdivided into two initiatives, namely Journal Descriptor Indexing (JDI) and Semantic Type Indexing (STI)[8-9]. Both initiatives have been developed based on manual classification of journals found in Medline database. And they are summarized below.

JDI methodology was designed to classify health-related texts based on statistical associations between words and 121 journal descriptors[10] while STI methodology[8] uses a group including 135 semantic types from Unified Medical Language System (UMLS) for text classification. Based on STI method, text classification is performed using similarity estimation[11] between vectors of journal descriptors linked to semantic types and texts.

Health Informatics Thesaurus

The Health Informatics Thesaurus (HIT), called EpistemIS[12-14], was created based on a theoretical platform with the purpose of reducing ambiguity of concepts, and conceptual relations, and promoting the development of specialized HI languages. It was constructed a hierarchical structure from surveyed HI concepts supported by a study on principles as proposed by Mario Bunge[15].

This thesaurus contains 730 terms, of which 620 were obtained from HI technical scientific literature. The remaining 110 terms were extracted from the *Medical Subject Headings* (MeSH), a specialized subject heading developed by NLM. MeSH includes health science terms also called descriptors. These descriptors are hierarchically organized and they allow searches at different levels of specificity.

The set of 730 terms is subdivided into the areas of Information Sciences, Biological Sciences, Behavior Sciences, and Human Sciences. EpistemIS thesaurus also provides epistemological classifications of terms and this classification is known as Metaconcepts of Action and Human Thought (MAHT)[12]. Table 1 shows some examples of terms and descriptions found in EpistemIS thesaurus.

Table 1 - Terms of a specialized HI thesaurus[12].

Term	Description
Health Informatics Computing	Precise procedural mathematical and logical operations utilized in the study of medical information pertaining to health care.
Societies, Health Informatics	Societies whose membership is limited to health informatics professionals and researchers.
Statistics	The science and art of collecting, summarizing, and analyzing data that are subject to random variation. The term is also applied to the data themselves and to the summarization of the data

Problem statement

Health Informatics research databases have dramatically increased the volume of discovery in the biomedical sciences. Medline summarizes knowledge that has been published across all biomedical fields. Medline and its search interface, PubMed, have concentrated efforts to retrieval of papers according to their subject content. Papers in Medline are indexed by MeSH.

A typical reason for constructing a controlled vocabulary like MeSH is to give a common language for sharing and reusing knowledge about phenomena in the world of interest, helping researchers to found information[16].

Authors believe that the Health Informatics is an interdisciplinary techno-science[13], it using knowledge corpus from other fields like Health and Computer

Sciences as well as Cognitive and Information Sciences. Thus, the main purpose of EpistemIS thesaurus is to start a discussion about a common language to represent HI through a hierarchy structure and grouping knowledge from Information Sciences, Biological Sciences, Behavior Sciences, and Human Sciences related to Health Informatics.

The experiment focus is to distinguish between Health and HI articles using well-know methods for classifying articles and EpistemIS thesaurus, then, recognizing the relevance of EpistemIS's descriptors to Health Informatics domain.

METHODS

The present study was conducted in three steps. In the first step, through an automated procedure, titles and abstracts of 900 HI scientific papers were collected. These articles were published between 2006 and 2009 and they were distributed uniformly in 9 journals. We also collected 900 articles from Health, published between 2004 and 2009 and distributed uniformly in 9 journals too. Both, HI and Health papers were collected according categories of ISI Web of Knowledge assigned to the journals. The papers are available at PubMed. The Table 2 below summarizes the both collections.

After, methods proposed by the TC project were adapted and then applied to the collected articles and EpistemIS thesaurus for estimation of the similarity coefficient between thesaurus terms and selected texts. The greatest similarity coefficients indicate greater relevance to HI's field.

Lastly, artificial intelligence methods, described later in this section, was also applied to the results obtained in the second step to check for different behaviors between specialized HI articles and all other articles based on the computed similarity coefficient.

The combined three steps described above were defined as Health Informatics Indexing (HII) in this work. The resources and methods used are detailed below. The flow chart in Figure 1 describes the sequence of HII methods applied.

Health Informatics Indexing (HII)

The HII initiative studies the application of statistical, vector, and artificial intelligence methods for classification of HI-related content.

In the present study, it was estimated the degree of similarity between 730 terms, found in EpistemIS, and collected articles, from HI and Health. Adapted methods from JDI and STI initiatives, as described in TC project, were used in the analysis.

Journal Descriptor Indexing (JDI) Details

The JDI methodology is based on words found in titles and abstracts of articles extracted from nearly 4,000 journals of Medline and their statistical associations with 121 journal descriptors (JD) from MeSH, which index the journals, e.g., 'Behavioral Sciences', 'Biomedical Engineering', 'Education', 'Health Services Research', 'Medical Informatics', 'Medicine', 'Nuclear Medicine', 'Radiology', 'Statistics'. Experts used these descriptors for manually classifying journals and this task can be considered minimal human effort compared to human indexing of individual articles. Thus, extracted articles "inherit" JD's from journal records corresponding to the journals in which the documents are published. Each word in the titles and abstracts of extracted articles can be said to co-occur with the JD by virtue of this inheritance[8].

The relevance of occurrence of journal descriptors for each word can be estimated in two different manners: 1) word-count method: is the sum of co-occurrences of

Table 2 - Distribution of categories, journals and papers collected.

Collection	ISI Web of Knowledge's category	# journals	# papers
Health Informatics	(1) Medical Informatics	9	900
Health	(1) Medicine, Research & Experimental; (2) Pediatrics; (3) Nursing; (4) Anatomy & Morphology; (5) Clinical Neurology; (6) Microbiology; (7) Biochemistry & Molecular Biology; (8) Oncology; (9) Biology	9	900

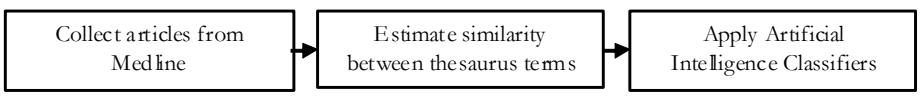

Figure 1 - Sequence of HII methods applied.

Sum of the number of times that the word *pulmonary* co-occurs with the journal descriptor *Surgery* in all articles.	= Score of relevance
Sum of the occurrence of the word *pulmonary* in all articles, regardless of the journal descriptor associated.	

Figure 2 - Example of the calculation of the word-count method for the descriptor *Surgery* and the word *pulmonary*.

a word to a journal descriptor, divided by the sum of occurrences of this same word in the articles, regardless of the descriptor associated; 2) document-count method: is the sum of articles showing the occurrence of a given word that co-occurs with a journal descriptor, divided by the sum of articles showing the occurrence of this same word, regardless of the journal descriptor associated.

Figure 2 and Figure 3 show the calculation of word-count and document-count methods for the journal descriptor *Surgery* and the word *pulmonary*.

Table 3 shows the distribution of the word *pulmonary* with regard to 5 journal descriptors. The columns Word and Document on the left represent the sum of co-occurrences of the word to the journal descriptors, meeting the restriction of word-count and document-count methods, respectively.

The relevance of occurrence of journal descriptors for the word *pulmonary*, obtained after word-count and document-count methods application, based on occurrence data displayed in Table 3, can be seen in this same table in columns Word and Document on the right. There is a remarkable difference in the results of these two methods; however, both show that the journal descriptor *Pulmonary Disease (Specialty)* has a higher score of relevance to the word *pulmonary*.

Once JD profiles have been computed for each word in titles and abstracts of articles extracted from journals, as described above, they can be used as the basis for indexing documents outside the 4000 Medline journals selected. We can assign JD's as indexing terms to a new document based on the words in the new document that also occur in the articles from 4000 Medline's journals selected and used to compute the word JD profiles, assuming the application of normalization methods such term frequency and the average of the rankings for JDs across all the words[17].

Similarity between thesaurus and collected articles

This step assigns JDs as indexing descriptors to EpistemIS terms and articles, from HI and Health, considering the JDI methodology and document-count approach described in the last section. The Table 4 shows matrix obtained after this assignment for EpistemIS, where the columns represent the EpistemIS terms ($Term_i$), the rows are the JDs(d_j), and content stored in element "a" is the relevance score of each JD for EpistemIS terms. A similar matrix was constructed for HI and Health articles; the only difference is the number of columns that are 900 for HI and Health, due to the number of articles in each collection, as showed in Table 5.

In this step it was also estimated the similarity between the vector of JDs of each EpistemIS term and vector of JDs of each article, from Health and HI. For this procedure it was applied the adapted vector space model as proposed by Salton and McGill[11], described in Equation 1. Table 6 shows similarity matrix created to represent the contextual usage of EpistemIS terms in relation to all articles, from HI and Health. In this matrix, the rows represent the 730 EpistemIS terms (t_i), the columns represent all articles ($Articles_j$), from HI and Health, and element "s" is the similarity value calculated according to Equation 1. Thus, after this process, the 730 EpistemIS terms are assigned to each HI and Health article sorted by the degree of similarity ranging between 0 and 1, where $TERM_i$ and $ARTICLE_j$ denote vectors of journal descriptors from EpistemIS terms and articles, respectively.

The last step provides the element 's', according Table 6, computed for each HI and Health article, as feature for training pattern classifiers. To reduce the dimensionality of the vector (column from Table 6) from 730 to 10, its values were distributed in 10 categories. Table 7 shows

Sum of articles where the word *pulmonary* co-occurs with the journal descriptor *Surgery*.	= Score of relevance
Sum of articles where the word *pulmonary* occurs, regardless of the journal descriptor associated.	

Figure 3 - Example of the calculation of document-count method for the descriptor *Surgery* and the word *pulmonary*.

Table 3 - Example of co-occurrence of the word *pulmonary* with regard to five journal descriptors.

	# occurrences		Scores of relevance	
Descriptors	**Word**	**Document**	**Word**	**Document**
Pulmonary Disease (Specialty)	2428	922	0.2645	0.2443
Cardiology	1326	401	0.1445	0.1063
Surgery	1215	460	0.1324	0.1219
Critical Care	499	199	0.0543	0.0527
Anesthesiology	368	151	0.0401	0.0400
Total occurrences	9177	3774		

Table 4 - Relevance scores matrix for EpistemIS terms.

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	$Term_{730}$
d_1	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{1,730}$
d_2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	$a_{2,730}$
d_{121}	$a_{121,1}$	$a_{121,2}$	$a_{121,3}$	$a_{121,4}$	$a_{121,730}$

Table 5 - Relevance scores matrix for Articles from Health and HI.

	$Article_1$	$Article_2$	$Article_3$	$Article_4$	$Article_{1800}$
d_1	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{1,1800}$
d_2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	$a_{2,1800}$
d_{121}	$a_{121,1}$	$a_{121,2}$	$a_{121,3}$	$a_{121,4}$	$a_{121,1800}$

$$COSINE(TERM_i, ARTICLE_j) = \frac{\sum_{k=1}^{t}\left(TERM_{ik} \bullet ARTICLE_{jk}\right)}{\sqrt{\sum_{k=1}^{t}\left(TERM_{ik}\right)^2 \bullet \sum_{k=1}^{t}\left(ARTICLE_{jk}\right)^2}}$$

Equation 1 - Adapted vector space model as proposed by Salton and McGill[11,17].

the criteria for each category. Thus, to execute the experiment the training pattern classifiers received 10 features, which improved the performance of the classifiers.

Table 6 - Similarity matrix for HI and Health articles in relation to EpistemIS terms.

	$Article_1$	$Article_2$	$Article_{1800}$
t_1	$s_{1,1}$	$s_{1,2}$	$s_{1,1800}$
t_2	$s_{2,1}$	$s_{2,2}$	$s_{2,1800}$
t_{730}	$s_{730,1}$	$s_{730,2}$	$s_{730,1800}$

Table 7 - Criteria to reduce the dimensionality of vector 's'.

Category	Criteria	Category	Criteria
1	$0 >= s_{i,j} >= 0.1$	6	$0.5 > s_{i,j} >= 0.6$
2	$0.1 > s_{i,j} >= 0.2$	7	$0.6 > s_{i,j} >= 0.7$
3	$0.2 > s_{i,j} >= 0.3$	8	$0.7 > s_{i,j} >= 0.8$
4	$0.3 > s_{i,j} >= 0.4$	9	$0.8 > s_{i,j} >= 0.9$
5	$0.4 > s_{i,j} >= 0.5$	10	$0.9 > s_{i,j} >= 1.0$

Text-Based Classifiers

In our experiments, six classifiers were used: Artificial Neural Networks (ANN)[18], Support Vector Machines (SVM)[19], K-nearest neighbours (KNN)[20], Naive Bayes[21], Bayes Net[22], Decision Trees. Different combinations of these classifiers' parameters[23] were tested to identify those with greater performance for classification of HI related content.

The classifiers was responsible to determine two classes, HI related content and not. To perform the experiments, the 10-fold cross validation method was adopted[24], where the dataset was randomly split in ten parts, thus, in each run, a different part was used as a test set while the remaining were used as training set. Conventional accuracy, precision, recall, F_1 measures and area under the ROC curve (AUC) were used to evaluate the performance of the presented method. Precision *p* is defined as the proportion of correctly classified examples in the set of all examples assigned to the target class. Recall *r* is defined as the proportion of correctly classified examples out of all the examples having the target class. F_1 is a combination of precision and recall defined as. ROC curve correlates sensitivity with the complement of specificity attained by any pattern recognizing tool for different ranges of values assumed by one or more parameters analyzed. Weka software program[25] was used as a tool for pattern recognition.

RESULTS

Figure 4 shows an example of HII method applied to the title and abstract of a text published in a HI journal collected. The right column shows the original MeSH descriptors assigned to the article and the two left columns show the EpistemIS terms assigned to the article and its similarity value in descendent order, obtained after applying the methods described in this article.

In addition, analysis of results was performed

Title: State-of-the-art anonymization of medical records using an iterative machine learning framework.

Abstract: OBJECTIVE: The anonymization of medical records is of great importance in the human life sciences because a de-identified text can be made publicly available for non-hospital researchers as well, to facilitate research on human diseases. Here the authors have developed a de-identification model that can successfully remove personal health information (PHI) from discharge records to make them conform to the guidelines of the Health Information Portability and Accountability Act. DESIGN: We introduce here a novel, machine learning-based iterative Named Entity Recognition approach intended for use on semi-structured documents like discharge records. Our method identifies PHI in several steps. First, it labels all entities whose tags can be inferred from the structure of the text and it then utilizes this information to find further PHI phrases in the flow text parts of the document. MEASUREMENTS: Following the standard evaluation method of the first Workshop on Challenges in Natural Language Processing for Clinical Data, we used token-level Precision, Recall and F(beta=1) measure metrics for evaluation. RESULTS: Our system achieved outstanding accuracy on the standard evaluation dataset of the de-identification challenge, with an F measure of 99.7534% for the best submitted model. CONCLUSION: We can say that our system is competitive with the current state-of-the-art solutions, while we describe here several techniques that can be beneficial in other tasks that need to handle structured documents such as clinical records.

MeSH	EpistemIS terms	Similarity Value
Medical Records Systems, Computerized	Clinical Decision Support Systems (CDSS)	0.9826
Humans	Performance of Medical Decision Models	0.9820
Evaluation Studies as Topic	Management of Decision Support Systems	0.9807
Confidentiality	Evaluation of Decision-Support Systems	0.9798
Artificial Intelligence	Classification Systems	0.9781

Figure 4 - List of MeSH and EpistemIS terms assigned to an article from HI corpus.

considering the percent distribution of similarity's scores assigned to articles. In this analysis, the scores of each article were distributed in tabulated frequencies, between 0 and 1, according Table 7. Figure 5 shows the percent histogram for the two groups of articles (HI and Health). Total distribution, as shown in Figure 5, reveals a significant difference between the two groups of articles. The curve representing similarity's scores of HI articles has greater concentration on the right side of the histogram. This behavior indicates that HI articles have higher similarity with EpistemIS than other articles.

Table 8 shows accuracy, precision, recall, AUC, and F1 measures, obtained as results of the evaluated classifiers. We consider F_1 measure the most representative value for our study; thus, the Artificial Neural Networks and Support Vector Machines classifiers provided the best results, although all classifiers presented very close results.

DISCUSSION

Relevance

The present study was based on previous studies for the application of methods for text classification and retrieval. Statistical, vector, and artificial intelligence methods were applied to obtain the results here presented.

The interdisciplinary of EpistemIS is enhanced through its association with the 121 MeSH descriptors, which reflect many areas used in the Health domain. The results obtained after estimating similarity between vector of JDs of each EpistemIS term and vector of JDs of each article show that the method adopted for this study was able to identify the presence of articles in the Health Informatics sub domain, which is part of Health domain. Figure 6 summarizes how the relationship between domain (Health) and sub domain (Health Informatics) was addressed in the article.

The authors believe that articles with HI content could be identified based on similarity with EpistemIS terms. A histogram was built with scores of similarity between EpistemIS terms and articles, from HI and Health, to check for visually distinct distributions between HI-related content and contents of other areas. Figure 5 shows a significant difference in the distribution of the two groups of articles studied, HI and Health. It is a positive distribution as the terms of the thesaurus studied, together with the proposed methods, were able to classify a higher percent of articles from journals specialized in HI, the area for which it was primarily designed.

The choice of using artificial intelligence methods to analyze histograms proved a valid approach as accurate results were obtained through the classifiers applied. The ANN and SVM classifiers were more effective when compared to others as it showed the highest performance for article classification, although it all classifiers' performance presented very close results.

The journals used in the JDI methodology construction cover different subjects in health and therefore did not use HI-specific descriptors for their classification. The proposed method shows that this is not an issue and it is possible, through similarity procedure, artificial intelligence techniques and HI-specific thesaurus, to classify HI related content.

Further research

1) The construction of a HI-specific training set based on a specialized thesaurus could significantly improve the

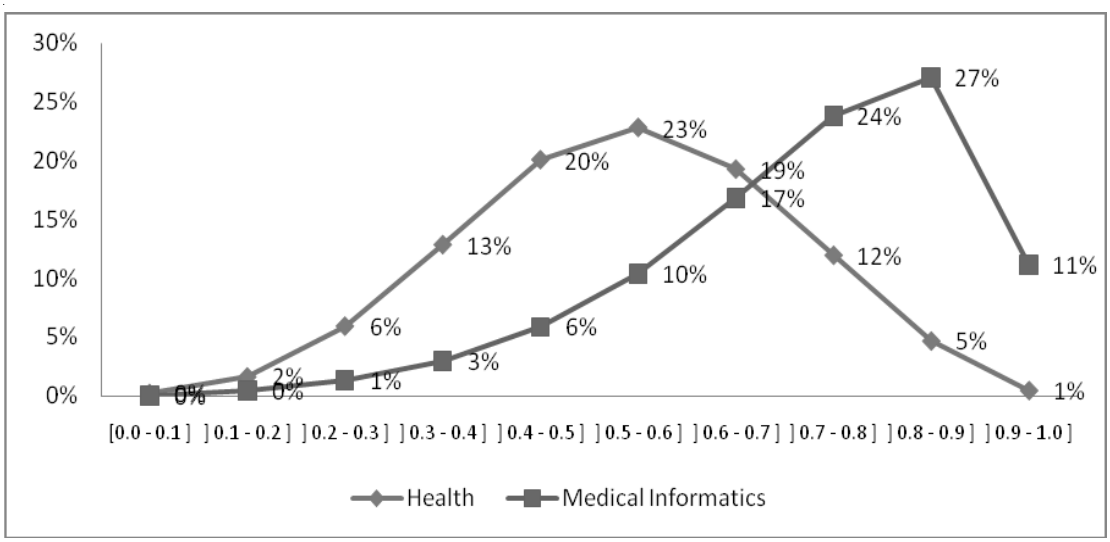

Figure 5 - Distribution of scores of similarity between EpistemIS terms and articles, from HI and Health.

Table 8 - Accuracy, precision, recall, AUC and F1 measures for classification of HI and Health contents, ordered by F_1.

Classifiers	Accuracy	Precision	Recall	AUC	F_1
Artificial Neural Networks	0.87	0.89	0.86	0.94	0.87
Support Vector Machines	0.87	0.89	0.86	0.87	0.87
K-nearest neighbours	0.86	0.88	0.84	0.93	0.86
Decision Trees	0.86	0.86	0.85	0.89	0.86
Bayes Net	0.85	0.88	0.82	0.91	0.85
Nayve Bayes	0.85	0.83	0.88	0.94	0.85

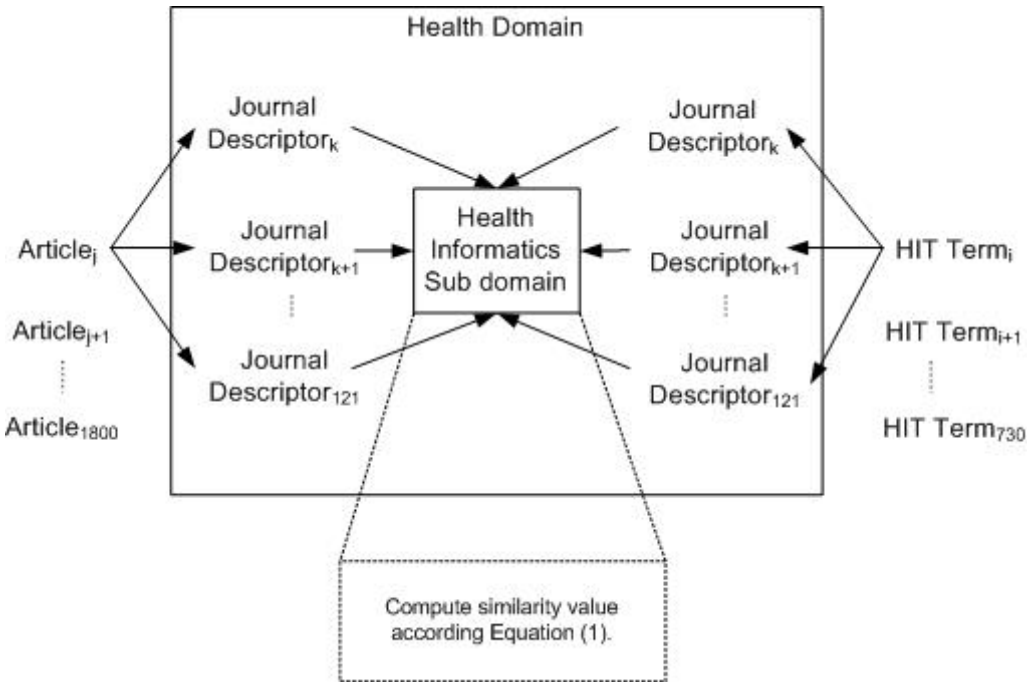

Figure 6 - Relationship between domain (Health) and sub domain (Health Informatics)

percentage of accuracy of the classification method investigated in the present study.

2) The similarity between two vectors can be estimated when they have the same number of components [10]. In our study we used 730 terms of EpistemIS to estimate similarity. This estimate can be simplified and provide better results if, supported by statistical studies, only those terms with the greatest score of similarity would be included.

3) Improve accuracy of concept-identification process, whereby a document is characterized as being associated with one or more concepts in the thesaurus.

4) Compare EpistemIS with others Controlled Vocabularies, like MeSH, and evaluate them using HII method and its capacity to classify HI-related content. Thus, we could identify what is the breakeven point between its descriptors.

CONCLUSION

Based on the results and evaluations of the present study, the application of the methods proposed by HII initiative can favorably contribute to the classification of HI-related content. The distribution of HI and Health articles (Figure 5) in the corpus of analysis with regard to their scores of similarity with EpistemIS terms presented distinct behavior after the application of the methods here described. Better results were achieved after applying artificial intelligence techniques. The F_1 measure and AUC, found in ANN classifier was 0.87 and 0.94, respectively. These are interim results in the study of methods and correlations between terms of a specialized HI thesaurus for potentially refining the classification and retrieval of articles within the area for which they were designed. We believe these methods can be further improved to minimize noise that may be affecting the accuracy of results.

We verified that the proposed methods can encourage the construction of automated tools for classification and retrieval of HI-related content.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the help given by the Health Informatics Department, Federal University of Sao Paulo. Funding: CAPES-REUNI.

REFERÊNCIAS

1. Gaudinat A, Ruch P, Joubert M, Uziel P, Strauss A, Thonnet M, et al. Health search engine with e-document analysis for reliable search results. Int J Med Inform. 2006;75(1):73-85.
2. Zanasi EA. Data Mining 8: Data, Text and Web Mining and Their Business Applications (Wit Transactions on Information and Communication Technologies). WIT Press; 2007.
3. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A Model for Evaluating Interface Terminologies. J Am Med Inform Assoc. 2008;15(1):65-76.
4. Lancaster FW. Vocabulary Control for Information Retrieval. 2nd ed. Info Resources Pr; 1986.
5. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: A model for author name disambiguation: Research Articles. J. Am. Soc. Inf. Sci.Technol. 2005;56(2):140-158.
6. Bolelli L, Ertekin ª, Giles CL. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation [Internet]. 31th European Conference on IR Research on Advances in Information Retrieval. Toulouse, France: Springer-Verlag; 2009 [cited 2009 Oct 3]. p. 776-780.Available from: http://portal.acm.org/citation.cfm?id=1533819
7. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. AMIA Symp. Philadelphia; 2000 Oct 17-21. EUA
8. Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best

semantic type based on Journal Descriptor Indexing: Preliminary experiment. J. Am. Soc. Inf. Sci. Technol. 2006;57(1):96-113.
9. Humphrey SM, Névéol A, Browne A, Gobeil J, Ruch P, Darmoni SJ. Comparing a rule-based versus statistical system for automatic categorization of MEDLINE documents according to biomedical specialty. J Am Soc Inf Sc Technol. 2009;60(12):2530-2539.
10. Lu CJ, Humphrey SM, Browne AC. A method for verifying a vector-based text classification system. AMIA Annu Symp Proc San Francisco, CA. 2008 Nov 10.
11. Salton G, McGill MJ. Introduction to Modern Information Retrieval [Internet]. McGraw-Hill, Inc. 1986 [cited 2009 Feb 3]. Available from: http://portal.acm.org/citation.cfm?id=576628
12. Colepicolo E, Novaes MAP, Pisa IT, Wainer J. Episthemology of the Medical Informatics [Internet]. 2005 [cited 2009 Feb 2]. Available from: http://www.icml9.org/program/poster5/
13. Colepicolo E. Epistemologia da Informática em Saúde: entre a teoria e a prática [Internet]. 2008;Available from: http://www.disacad.unifesp.br/sapg/arquivos/arq_55.pdf
14. Teixeira F. EpistemIS-WEB [Internet]. 2008. Available from: http://telemedicina6.unifesp.br/projeto/teixeif/BuscaDinamicaGraphSaude/
15. Bunge M. La Investigacion Cientifica: su estrategia y su filosofía. 3rd ed. Mexico: Siglo XXI; 2004.
16. Holsapple C, Joshi K. A collaborative approach to ontology design. Commun. ACM. 2002;45(2):47.
17. Humphrey SM, Rindflesch TC, Aronson AR. Automatic Indexing by Discipline and High-Level Categories: Methodology and Potential Applications.11th Asist Sig/Cr Classification Research Workshop. 2001 nov; Whashington, DC, USA; 2001.
18. Haykin SS. Neural networks and learning machines. Prentice Hall; 2008.
19. Chang C, Lin C. LIBSVM: a library for support vector machines [Internet]. 2001 [cited 2009 Oct 1]. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm
20. Aha DW, Kibler D, Albert MK. Instance-Based Learning Algorithms. Machine Learning. 1991;6(1):37-66.
21. John G, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. Eleventh Conference On Uncertainty In Artificial Intelligence. Montreal.Quebec, Canada.1995 Aug.
22. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992;9(4):309-47.
23. Teixeira F. Weka Classifiers for the article "Similarity-based scoring method for classification of Health Informatics content" [Internet]. 2008;Available from: http://telemedicina6.unifesp.br/projeto/teixeif/artigo/parameters.html
24. Burnham KP, Anderson D. Model Selection and Multi-Model Inference. 2nd ed. Springer; 2003.
25. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco, California, USA: Morgan Kaufmann; 2005.