PERGAMON

Information Processing and Management 39 (2003) 825–851

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Experiments in discourse analysis impact on information classification and retrieval algorithms

J. Morato *, J. Llorens, G. Genova, J.A. Moreiro

Department of Computer Science, Universidad Carlos III de Madrid, Av. Universidad, 30-28911 Leganés, Madrid, Spain

Received 1 February 2002; accepted 13 August 2002

Abstract

Researchers in indexing and retrieval systems have been advocating the inclusion of more *contextual information* to improve results. The proliferation of full-text databases and advances in computer storage capacity have made it possible to carry out text analysis by means of linguistic and extra-linguistic knowledge. Since the mid 80s, research has tended to pay more attention to *context*, giving discourse analysis a more central role. The research presented in this paper aims to check whether discourse variables have an impact on modern information retrieval and classification algorithms. In order to evaluate this hypothesis, a functional framework for information analysis in an automated environment has been proposed, where the *n*-grams (filtering) and the *k*-means and Chen's classification algorithms have been tested against sub-collections of documents based on the following discourse variables: "*Genre*", "*Register*", "*Domain terminology*", and "*Document structure*". The results obtained with the algorithms for the different sub-collections were compared to the MeSH information structure. These demonstrate that *n*-grams does not appear to have a clear dependence on discourse variables, though the *k*-means classification algorithm does, but only on domain terminology and document structure, and finally Chen's algorithm has a clear dependence on all of the discourse variables. This information could be used to design better classification algorithms, where discourse variables should be taken into account. Other minor conclusions drawn from these results are also presented.
© 2002 Elsevier Ltd. All rights reserved.

Keywords: Discourse-model; Context-analysis; Computational-linguistics; Text-analysis-methods; Filtering; *n*-grams; *k*-means; Co-wording

* Corresponding author. Tel.: +34-91-6249107.
E-mail address: jorge@ie.inf.uc3m.es (J. Morato).

0306-4573/03/$ - see front matter © 2002 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4573(02)00081-X

1. Introduction

Since Garfield (1953) developed his work in the fifties, linguistic analysis has always been related to the improvement of information tools. Nowadays, morphologic, syntactic and semantic analysis are included in commercial information retrieval (IR) systems (Warner, 1994), but contextual approaches are scarce.

At the beginning of the seventies, the interest of linguistics for improving text analysis stressed the importance of *context* (Dijk, 1988). Nowadays, this trend has an increasing influence in other areas such as automatic indexing and filtering (Llorens, Velasco, Morato, & Moreiro, 1998) or automatic translation. In this approach, the full understanding of a sentence requires that some words, like pronouns or adverbs, be interpreted in relation to the other sentences, in order to resolve anaphoric situations (that is, when the pronoun or adverb points to another term that has been already mentioned in the discourse). Some studies show that anaphoric references have a direct impact on the performance of natural language processing (NLP) tools. This is the case of pronoun analysis in automatic translation in Mitkov (1998) or in information classification algorithms in Llorens et al. (1998).

Although the pure text analysis approach is still needed to understand how texts are informationally, rhetorically and stylistically (Swales, 1990) organized, text analysis is insufficient for a holistic approach in IR systems. Certain features, like citing practices in academic papers, can only be detected with extra-textual information. A solution appeared in the 70s with the first works on the automatic analysis of discourse (Pêcheux, 1969). Discourse operates within conventions defined by academic disciplines and social groups. These social groups share some specific lexical items, linguistic forms, regulative rules and cultural concepts. It is in this context that some elements of discourse analysis, such as textual structures, genres and registers are employed to complete the context analysis (Karlgren & Cutting, 1994). In this paper, they are all examined in order to study their relative importance in IR tools.

In Section 2, a theoretical overview about discourse terminology is presented. The remainder of this article presents a methodology to study the impact of discourse aspects in IR systems.

2. A theoretical overview of some aspects of discourse

In this work, we consider discourse as an instance of language use whose type can be classified on the basis of such factors as genre, register, domain terminology, or document structure. Discourse analysis studies the organization of language above the sentence or paragraph, and therefore takes into consideration larger linguistic units.

Discourse analysis is a vast, yet little defined area of linguistics. One reason why this is so is that the concept of discourse is based on different approaches from a number of academic disciplines (Schriffrin, 1994; Beghtol, 2001). Although many professionals in linguistics consider discourse terminology to be ambiguous and confusing, two perspectives have in general been adopted for looking at discourse analysis: structural and functional. The structural perspective works with text, discovering regularities and analyzing units. The functional perspective works with context and style, explaining language in relation to its social function. In this paper, a holistic approach, both structural and functional, is proposed for discourse analysis.

A theoretical overview of some discourse aspects is presented in the next sub-sections. These aspects comprise both of structural and functional analyses. Four aspects will be discussed: genre, register, domain terminology and document structure. All of them seem to be strongly interrelated.

2.1. Genre

Genre could broadly be defined as "*a collection of communicative events that share a set of communicative purposes*" (Swales, 1990). However, there is no complete agreement among the scientific community on this definition: some authors consider *genre* to be part of the concept of register (Amitay, 1998), (described later on). Communicative purposes are identified and mutually understood by the members of the professional or academic community in which they regularly appear. Types of genres are: news broadcasts, recipes, press conferences, encyclicals, and so on. Indeed, researchers have carried out many Internet genre studies over the last few years. Other experiments have applied a discriminant analysis with several parameters to differentiate among genres (Karlgren & Cutting, 1994; Morato, 1999).

Within genre, the concept of *document typology* has a more specific meaning. Examples of document typologies are research and work-in-progress notes, research articles, conference proceedings, and so on. Document typology is widely employed by library and information science (LIS) researchers. LIS researchers have been studying the automatic classification of these typologies for a long time. Their aim is to increase the precision-recall and pertinence ratios. For example, Gilyarevsky, Uzilevsky, and Moudrov (1997) has studied automatic classification by means of title length of the articles in agriculture journals. Haas, Sugarman, and Tibbo (1996) has also developed a text filter by means of an automatic classification of the characteristic vocabulary in empirical articles. In this work, a similarity measure is calculated between two clusters: one of the clusters is built with empirical vocabulary extracted from documents, and the other cluster with the vocabulary from the documents to be tested. In this paper we are going to use genre to reference document typology.

2.2. Register

In order to define register, it is necessary to start with the definition of style. Karlgren (1998) defined style as the set of choices between different lexical structures, morpho-syntactic structures, and linguistic markers among documents dealing with a certain topic. Therefore, the style in language arises from the possibility of choosing from alternative forms of expression that are characteristic of a particular person, group of people, or period of time. Stylistic variations are found due to several factors. The main ones are: first of all, the intended audience and the discourse environment where the text is produced, and, secondly, the author's preferences and personal idiosyncrasy. Thus, style has a straight relationship with Goffman's hypothesis about regions in the Zipf's curve (Egghe & Roussau, 1990) characterized by low occurrence words.

Researchers in socio-linguistics used to describe style as register. Register is a contextual aspect that correlates the groupings of linguistic features with recurrent situational features (Halliday, 1985). Register represents more generalized stylistic choices. According to Halliday, human language is based on three main types of functions: the field, the tenor and the mode of discourse.

Although these three functions are related to register, tenor is the most important one because it predicts the selection of options in the interpersonal component (Lavid, 1995).

Register is strongly interconnected with other discourse aspects. For example, differences in language use, known as a sub-language, may arise because of the nature of the material, the discipline or the register. Losee (1996), has studied the grammatical characteristics of compound words and phrases in different disciplines to determine the sub-language being used.

2.3. Domain terminology

In the context of this paper, the definition of "domain" is closely related to "field of discourse". Its origin can be found in the software engineering field, and intends to define the scope where a set of software applications are applied. For example, the Banking domain refers to the representation of all the information needed to understand and develop banking software applications. "Domain terminology" refers to the set of terms that best describe a particular field of knowledge. A well-established domain terminology about a particular field depends on the maturity level of the domain as well as the human efforts to group the different terms in a vocabulary. Different works, particularly those coming from the software/knowledge engineering arena, conclude that domain terminology has a strong impact on the performance of computer-based algorithms. Prieto-Díaz (1988) advices to apply domain analysis (DA) techniques only to mature domains.

Domain terminology is also expected to produce a strong impact in automatic domain construction, this deals with the identification of concepts (terms) describing the domain and relationships between them. This technology is usually known as DA. Several experiments have been performed over the last couple of years to build domains automatically (Neighbors, 1981; Díaz, Velasco, Llorens, & Martinez, 1998). Llorens, Velasco, and Martínez Orga (1997), have applied this technology to the automatic generation of thesauri. The method works by identifying meaningful terms by means of a set of text filtering techniques (Frakes & Baeza-Yates, 1992). The selected terms are usually the most representative of the document sets. Then an algorithm clusters the terms to establish their hierarchical and horizontal relationships. Polanco, Grivel, and Royauté (1995) has also used a clustering methodology to identify bibliometric variables applied to the diachronic study of terminological variations. Callon, Courtial, and Penan (1993) approach studies the evolution of research trends with cluster analysis. The results obtained by de Looze and LeMarié (1997) in Co-wording states that in order to get a detailed and complete image of a domain it is necessary to consult a fairly large number of databases and to analyze several corpuses.

2.4. Document structure

In order to assign overall semantics to a set of words, they must have an underlying linguistic structure throughout the text (Leydesdorff, 1997) usually called document structure. Documents belonging to particular genres, such as research papers, are often highly structured and conventionalized with constraints. Several authors (Dijk, 1988; Hearst & Plaunt, 1993) suggest linking the description keywords or terms representing the document index with the document structure where they occur. Kando (1997) has also shown that using text-level structures in searching

achieves higher precision rates in IR systems. More specifically, Leydesdorff (1997) has claimed that the patterns of co-absences and co-occurrences are specific for each section.

The precise lay out of the information within the text structure can be a valuable factor. Cognitive experiments show that common strategies to detect worthwhile information are rarely accomplished by reading articles sequentially. Some studies (Swales, 1990) show that there is a tendency to look at the abstract first then at the conclusion, followed by the figures and tables, and finally the results.

3. The experimental framework for studying the impact of discourse variables in indexing and classification algorithms

In this section, we present a methodology and experimental framework to evaluate the impact of discourse variables within information science techniques. The main aim is to provide a means of evaluating the impact of different genres, registers, domain terminologies or document sections in two different types of algorithms: indexing and classification. The first step of the methodology (presented in Fig. 1) is to build a collection of electronic documents gathered from document databases. These documents must be pre-processed in order to get a uniform structure for all of them. The collection is divided into sub-collections based on different discourse variables and indexing and classification algorithms are later applied to all the sub-collections in order to contrast the results for different genres, registers, domain terminologies and document structures. The results obtained from the application of the algorithms to the discourse-based sub-collections must be compared to a well defined accepted information structure in order to evaluate the impact of the different discourse variables in the algorithms: the medical subject headings (*MeSH*) thesaurus. The evaluation criteria was: "the closer the results of the algorithms come to the MeSH

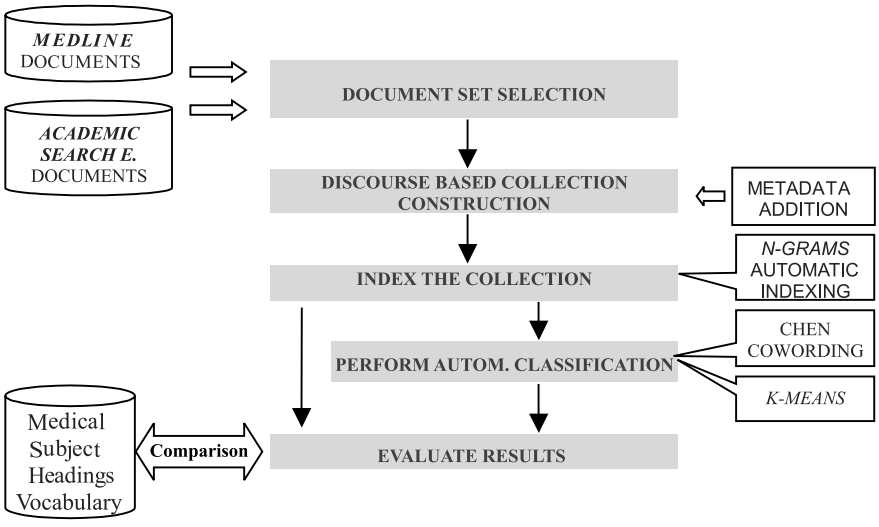


Fig. 1. Methodology scheme.

structure, the better the algorithms have performed''. The reason for applying this criteria was that the MeSH associations were widely accepted among medical experts, as discussed in Section 4. On the other hand, it was also thought interesting to include some comparisons with classical IR algorithms in order to evaluate them alongside MeSH (i.e. see Section 4.1). The absolute results (which variable performs best), and specially the relative results (comparison of results for different values of the variables within the sub-collections) allow us to measure the impact of the different discourse variables in the algorithms studied''. The methodology is shown in the next sections:

3.1. Document set selection

The selected collection of documents consisted of 440 full-text electronic documents. The documents were extracted from two different databases: *Medline* and *Academic Search Elite*. These databases were selected for their accessibility and the presence of full-text articles from prestigious publications.

The collection was selected to cover five different genres (''Research articles'', ''News'', ''Conference proceedings'', ''Notes'', and ''Popular-Science articles''), three different registers (''Scientific language'', ''Press language'', and ''Popular-Science language''), five different domain terminologies (''Hepatitis'', ''HIV'', ''CJD'', ''Botanical Proteins'', and ''Clinical Proteins'') taken from two different domains, medicine and biology, selected because their organization and vocabulary are highly normalized (Nwogu, 1997) and six different document sections (''Abstract'', ''Introduction'', ''Methods'', ''Results'', ''Discussion'', ''References'').

The following publications, from 1996, were selected from the Medline database:

- New England Journal of Medicine
- British Medical Journal, Lancet
- Journal of Clinical Investigation
- AIDS Care

In Academic Search, the following genre-based publications were retrieved:
For news:

- US News & World Report
- The Economist
- Newsweek
- Time

For research articles:

- Plant Physiology

and for popular-science articles:

- Blood Weekly

3.2. Discourse-based sub-collection construction

In order to analyze the impact of the different discourse variables in the algorithms a subdivision criterion was applied to document collection, based on one of the following four discourse variables: genre, register, domain terminology, or document structure. Four different groups of sub-collections were created one for each discourse variable. Each group included the same 424 documents (apart from the document structure group that only had 94 out of the 424) but grouped into sub-collections according to the value of the discourse variable. For example, in the genre group, the 424 documents were separated into five sub-collections, where all the research articles were placed together in the same sub-collection, all the News in another sub-collection, and so on. The documents were manually assigned to the collections. The result was the following 19 sub-collections, with homogeneous documents in each one:

- Genre group (five sub-collections): "Research articles", "News", "Conference proceedings", "Notes", and "Popular-Science articles".
- Register group (three sub-collections): "Scientific language", "Press language", and "Popular-Science language".
- Domain terminologies group (five sub-collections): "Hepatitis", "HIV", "CJD", "Botanical Proteins", and "Clinical Proteins".
- Document structures group (six sub-collections): "Abstract", "Introduction", "Methods", "Results", "Discussion", "References".

3.2.1. Genre sub-collections

The creation of the five genre sub-collections was based on the document typology and the Journal they came from. Notes, and conference proceedings were gathered from Medline, from where news and popular-science articles were selected. Academic Search Elite Research articles were found in both Medline and Academic Search Elite. Research papers should have had scientific characteristics and should have been evaluated by independent referees. Short papers, not showing scientific characteristics but found in scientific publications were considered as Notes. Documents were assigned to news genre when they had a high novelty content and did not describe the concepts presented in the articles. Several documents from newspapers (US News, etc.) were manually assigned to popular-science when they met the following criteria: (1) Their aim was to explain a particular scientific subject from the very basics to an unskilled reader. (2) The paper described information gathered as a result of long term research, where its impact in the science field is expected to last. Editorials were not considered in this variable.

The sub-collection structure is given in Table 1.

3.2.2. Register sub-collections

All the documents were read and assigned to different register variables in a subjective manner in order to create the Register sub-collections. The following results were gathered; all the research articles were considered to contain "Scientific language". "Conference proceedings" and "Notes" were almost all assigned to scientific language, the rest were assigned to "Popular-Science". One part of News articles was assigned to "Popular-Science" and the rest to "Press language". Thus, a clear correspondence cannot be deduced for genre and register results. The

Table 1
Genre sub-collections

	Sub-collections					
	Research articles	News	Conference proceedings	Notes	Popular-science articles	Editorial
New England Journal of Medicine	23			8		6
British Medical Journal	26			35		1
Journal of Clinical Investigation	48					
Lancet	35			46		7
AIDS Care	43		1	5		
US News & World Report		15			1	
The Economist		11			1	1
Newsweek		11		1	1	1
Time		6			2	
Plant Physiology	44					
Blood Weekly			9		36	

Table 2
Register sub-collection

	Sub-collections		
	Scientific language	Press language	Popular-science language
New England Journal of Medicine	37		
British Medical Journal	62		
Journal of Clinical Investigation	48		
Lancet	88		
AIDS Care	49		
US News & World Report		16	
The Economist		9	4
Newsweek		12	2
Time		7	1
Plant Physiology	44		
Blood Weekly			45

documents coming from newspapers whose main aim was to explain scientific terminology were considered to contain popular-science language. Blood Weekly papers were considered as Popular-science language as they followed the features defined by Posterguillo (1996) to identify them: (reduction of locations, etc.).

The sub-collection structure is given in Table 2.

3.2.3. Domain terminologies sub-collections

Nearly all the publications from Medline belong to the 'Medicine General and Internal' JCR thematic group. The Journal of Clinical Investigation, which belongs to the 'Medicine Research Experimental' thematic group, was selected in order to introduce domain terminology differences in the collection.

Table 3
Domain sub-collection

	Sub-collections				
	Hepatitis	HIV	CJD	Botanical proteins	Clinical proteins
New England Journal of Medicine	12	24			1
British Medical Journal	13	40	9		
Journal of Clinical Investigation		1			47
Lancet	22	65	1		
AIDS Care		49			
US News & World Report		16			
The Economist		12	1		
Newsweek		14			
Time		8			
Plant Physiology				44	
Blood Weekly	17	28			

In the medicine domain, the document selection criterion consisted of three pandemic diseases: Hepatitis, AIDS, and Creutzfeldt–Jakob disease (CJD), which were selected to compare different research stages. Hepatitis is a spottily distributed disease of great antiquity in medical literature whose vocabulary is quite normalized. AIDS was a recent and increasingly global disease in 1996 (when this work started to be designed), and CJD was also a sporadic and infrequent emerging disease in rural areas in 1996, and its vocabulary is at a primary stage.

In the biology domain, 91 documents were retrieved from the botanical and biochemistry disciplines. The different domain terminologies selected were botanical proteins from Plant Physiology Journal, and clinical proteins in the Journal of Clinical Investigation.

The sub-collection structure was given in Table 3.

3.2.4. Document structure sub-collections

In order to study if indexing and classification algorithms perform in a different way depending on the document section they work with, 94 of the total 424 documents were structured in the IMRD organization (Introduction, Method, Results and discussion) proposed by Bruce (Bruce, 1983). 94 scientific articles were structured according to an expanded IMRD structure including Abstract, and References. When the selected articles did not include a particular section, it was added as an empty one. The rest of the documents, particularly those coming from the ''News'' genre, were not considered due to difficulties in structuring them using the expanding IMRD structure. The assignation process was performed automatically using a computer-based matching program. The algorithm matched single terms located in the different section titles of the documents with a term list for every different expanded IMRD section. The highest success rate was found in the Abstract, Introduction and References sections. However, the matching process was fully controlled by the authors in order to recover failures. Although the IMRD organization is widely used in medicine and biology, a significant percentage of the 94 documents did not follow this structure originally. In order to transform the different sections of the documents into the IMRD organization, the authors used the previous experiences of Swales (1990), Nwogu (1997) and Skelton (1994).

Table 4
Document structure sub-collection

	Sub-collections					
	Abstract	Introduction	Methods	Results	Discussion	References
New England Journal of Medicine	20	20	15	14	17	20
British Medical Journal	22	22	22	22	22	21
Lancet	28	28	28	28	28	28
AIDS Care	24	24	24	24	21	20

The result of this pre-processing was a collection of documents with an expanded IMRD structure (Table 4).

3.3. Testing indexing-retrieval algorithms

In classical IR, one of the most important tasks deals with the selection of relevant terms describing the target documents. According to this idea, when an electronic document has to be treated, computer programs try to automatically select which terms from the full text must be selected as document "descriptors". This process is usually made by the *n*-grams algorithms (Cohen, 1995), stop-word removal, tf–idf algorithms (Spark Jones, 1972), or simply human removal and it is usually known as text-filtering (Frakes & Baeza, 1990). Both *n*-grams or tf–idf are algorithms that use frequency information from every document as well as inter-document information. In our study, the *n*-grams algorithm has been selected to test the impact of discourse variables.

3.3.1. n-grams term-filtering algorithm

n-grams performs term-filtering by splitting the input text into grams (or groups) of *n* characters. For example: Using a 5-grams algorithm the "carbon monoxide" input text should generate the following 5-grams, "carbo", "arbon", "rbon", "bon m", "on mo", etc.

These grams are statistically compared against a set of grams from a background collection. The filtering process is made by accepting those terms that include the accepted grams. The best advantage of this algorithm as an indexer is that it is capable to select as descriptors (text filtering) not only single terms but noun phrases. However, decisions about gram size, as well as size, discipline and detail of the background lead to different results (Díaz et al., 1998).

To calculate the relevance of every gram the following formula is used:

$$y_i = \begin{cases} C_i \ln(C_i/S) + B_i \ln(B_i/R) - (SC_i + RB_i) \ln[(SC_i + RB_i)/(S + R)], & SC_i \geqslant RB_i \\ 0, & SC_i < RB_i \end{cases}$$

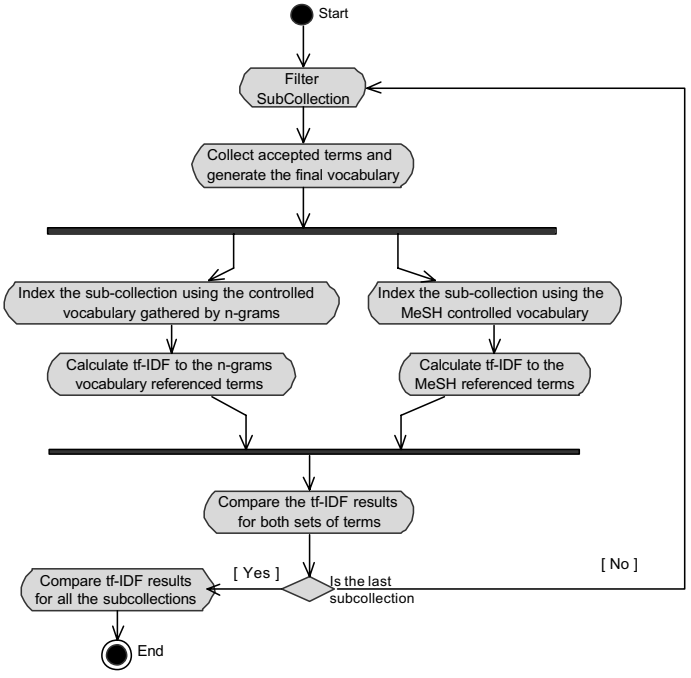
where C_i represents the value of the *i n*-gram in the document, B_i represents the value of the *i n*-gram in the background, S represents the value for the set of *n*-grams in the document, and R represents the value for the set of *n*-grams in the background. Several design decisions must be considered when using *n*-grams (Llorens et al., 1998). One of these decisions is how the collection

is processed. *n*-grams discriminate among the descriptors, accepting only those with intermediate appearance frequencies. These intermediate frequencies represent the concepts of the document better than descriptors with higher or lower frequency, as is predicted in the Zipf distribution. Nevertheless, words that appear only once in a document could represent important information. Another decision taken was the size of the *n*-grams. Our experiments have shown that odd values, from three to seven characters for each gram, provided better descriptors.

As *n*-grams can work as a term filtering algorithm, it can also be applied to generate a candidate set of terms which should represent a domain, from an electronic document corpus. This feature will be used to contrast the impact of different discourse variables in the algorithm, by measuring the quality of the candidate corpus when it is compared to MeSH.

3.3.2. Test framework for n-grams

The structure of the experiment to test the *n*-grams algorithm is presented by the following diagram:



For every sub-collection

It must be filtered using *n*-grams.
A term vocabulary is created from the terms selected by *n*-grams.

Two activities are performed in parallel:

(a) Index and calculate tf–idf for the vocabulary gathered by *n*-grams.
(b) Index and calculate tf–idf for the MeSH vocabulary.

Compare intra-sub-collection results (to check absolute performance)

After treating all the sub-collections

Compare inter-sub-collection results (to measure the impact of discourse variables in the algorithm)

Once the filtering algorithm selected the relevant terms of the collection, the set of documents was referenced. Two index references were made: one using the *n*-grams vocabulary as controlled vocabulary and the other one using MeSH vocabulary. The main purpose of this process was also to verify the quality of *n*-grams algorithms in different discourse variables. tf–idf was used to compare both indexing processes. The basic principle of idf is that the importance of a term within a document is higher when its frequency among all the documents is low. In IR, a high tf–idf value for a term implies that it must be selected to create the document index. In the indexing process, each descriptor was tabulated in the database, together with the number of its occurrences and the section of the paper where the term appears. A more detailed description of this process can be found in Díaz et al. (1998).

3.4. Testing classification algorithms

One of the main research activities in IS regarding information organization deals with developing algorithms that automatically find relationships among text terms. These algorithms are usually called classification algorithms. In order to test them, a comparison between the relationships created by the algorithms and a well established term relationships structure, as MeSH, will be made.

Two well-known algorithms were selected to obtain the relationships among terms: *k*-means and Chen's algorithm. These two methods were chosen because they provided good results in previous works (Díaz et al., 1998).

3.4.1. k-means classification algorithm

k-means is one of the most popular clustering techniques. It has usually been employed to generate clusters of objects with common features. In information science, these objects can be documents, users, references, queries, or, as in our study, clusters of terms found in the sub-collections.

This algorithm was used to accomplish a top–down approach, which facilitates the identification of hierarchies for the domain representation. *k*-means belongs to the family of "moving center" cluster analysis algorithms (Lelu, 1993). This means that the centroid of a group of terms is recalculated after a new set of document terms is inserted. *k*-means involve a number of critical input parameters, which are used to control the classification process, such as the number of desired clusters or the criteria established to select the descriptor that forms the root in the hierarchy.

3.4.2. Test framework for k-means

In our framework, *k*-means must be computed after the indexing process. The construction of hierarchies is done using a top–down approach. First, a root must be selected; after selecting the root, a clustering process must be done with the rest of descriptors using *k*-means. The input for *k*-means is a set vectors, one for each term from the vocabulary gathered in the *n*-grams filtering

process. Every component for a vector represents the relative number of occurrences for the term in a particular document. The application of *k*-means to the set of vectors creates different clusters, as well as information about the centroid for every cluster.

When the process was finished, the different clusters were treated as specific terms of previous root. Using the extraction of principal components over each cluster it was possible to obtain the next level in the hierarchy. These new roots were specific terms over first level roots. The methods used to extract roots were the "centroid distance", "large number of occurrences", "large number of documents" and "generality coefficient" (Díaz et al., 1998).

In order to test discourse effect in *k*-means affectivity, the cluster results gathered by *k*-means must be compared with the MeSH-tree hierarchies. One critical problem is that *k*-means needs an input parameter, which indicates the number of clusters desired. Estimating this value is a critical decision. If the value is low, the algorithm will create few clusters and therefore, the probability for every cluster to include MeSH hierarchies will be very high. On the other hand, a high value of clusters will imply the coverage of fewer hierarchies from MeSH. This value can be equaled to the number of clusters found in MeSH. Therefore, the experiment must first identify hierarchies in MeSH (Fig. 2).

As *k*-means does not provide directly hierarchies, but clustered terms, all the terms in every MeSH hierarchy will be grouped in a cluster and then compared with the *k*-means clusters (Fig. 3).

3.4.3. Chen–Co-wording classification algorithm

Chen's algorithm is a variation of the Co-wording algorithm. The Co-wording algorithm arose from different proposals related to bibliometrics. It basically consists in building up science maps by means of extracting associations from word occurrences (Callon, Courtial, Turner, & Bauin,

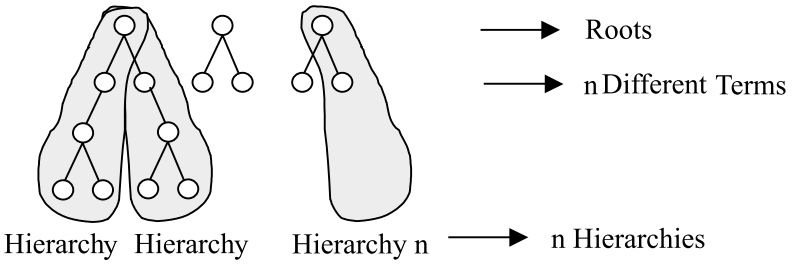


Fig. 2. Selection of hierarchies in MeSH.

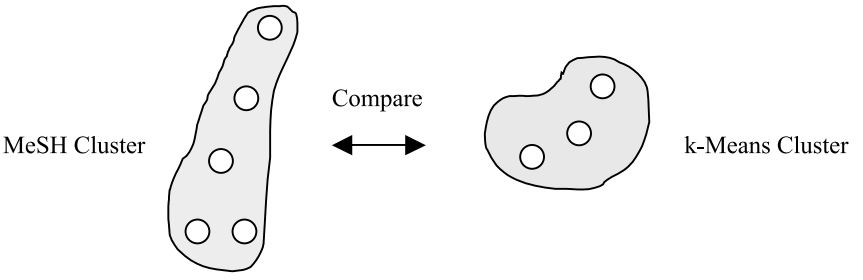


Fig. 3. Test framework for *k*-means.

1983). The Co-wording essential hypotheses is that two terms are semantically related when they usually occur jointly in the same document. Thus, it is possible to measure the semantic distance between two terms by means of computing the co-occurrences and co-absences in the collection.

The Chen method works with Co-wording. This algorithm generates a coefficient, for each pair of terms, that measures the degree of relationship between them (usually association). The result of this algorithm is a matrix of term relationships.

A coefficient is established for each pair of terms to indicate the cluster weight.

$$\text{Clusterweight}(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \text{Weightfactor}(T_k)$$

$$\text{Clusterweight}(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ik}} \text{Weightfactor}(T_j)$$

where d_{ij} and d_{ik} is tf–idf and $d_{ijk} = \text{tf}_{ijk} \log(N/\text{df}_{jk})$ is the tf–idf value when the terms j and k are in the same document i (Co-wording), thus, tf_{ijk} is the number of occurrences of both term j and k in document I, and df_{jk} represents the number of documents with the terms j and k.

A coefficient is established for each pair of terms to indicate the cluster weight (a number that belongs to the interval (−1,1)) between them. This coefficient is compared to a threshold. According to the value of the threshold the type of relationship between the descriptors is inferred. Taking the descriptors set and the values that have surpassed the threshold a graph can be built.

Chen's technique only intends to establish relationships between terms but does not differentiate between each type of relationship (equivalence, permanent association); the technique only works for a global type of relationships: associations.

3.4.4. Test framework for Chen/Co-wording

In our project, the *Chen algorithm* was applied after the indexing process. A coefficient was established for each pair of terms to indicate its association level. The weight was calculated by computing the inverse document frequency (idf) and the within-document frequency (tf) (Chen & Lynch, 1992).

The structure of the experiment was the following:

The terms set generated in *n*-grams was used as the input to the Chen algorithm. The result was a group of binary associations between terms. The associations were checked against the existing associations in MeSH.

4. Experiments and results

Three main experiments were performed to study the impact of discourse variables in IR algorithms:

Experiment 1:

Evaluate the dependence of *n*-grams algorithm on discourse variables. Run *n*-grams filtering algorithm, in different discourse-based sub-collections, with the intention of creating a set of terms that could form a domain vocabulary for every sub-collection. Then, index the document sub-collections first using the *n*-grams terms as controlled vocabulary and then the MeSH

vocabulary. Calculate tf–idf values for all the referenced terms, and compare the frequency results for both controlled vocabularies on each discourse variable sub-collection.

Aims:

(1) Measure the quality of *n*-grams as a filtering and vocabulary generation algorithm.
(2) Evaluate the impact of discourse variables in the *n*-grams algorithm.
(3) Study the dependence of domain terminology from the other discourse variables: in particular, to study how close domains use different synonyms to describe the same concept.

Experiment 2:

Evaluate the dependence of *k*-means algorithm to discourse variables. Run *k*-means classification algorithm for the terms gathered in the different discourse-based sub-collections, in order to create term-based clusters and compare the results with the MeSH tree hierarchies.

Aims:

(1) Measure the quality of *k*-means as a hierarchies-oriented classification algorithm.
(2) Evaluate the impact of discourse variables in the *k*-means algorithm.

Experiment 3:

Evaluate the dependence of the Chen algorithm on discourse variables. Run the Chen association algorithm in different discourse-based sub-collections, to create associations between terms and compare the results with the MeSH tree.

Aims:

(1) Measure the quality of Chen as an associations-oriented classification algorithm.
(2) Evaluate the impact of discourse variables in the Chen algorithm.

The authors consider that one of the most interesting properties of the proposed framework is that it is very independent from the selected algorithms. As the main aim of the framework is to compare the performance of the algorithms when they are applied to discourse-based collections of documents, the most interesting conclusions can be gathered from the relative comparisons of results.

In order to enhance the indexing and classification processes, specific information was stored in a database that should be used by the indexers and classification algorithms:

- A stop list from the *SMART project* (SMART, 2001) was tabulated. Before indexing, every stop-word was removed in accordance with this list.
- An English dictionary was added to the database. Each entry was tagged with its grammatical category. The source of this dictionary was WordNet and the *British National Corpus* (BNC, 2001).
- The well-known MeSH was incorporated in the database. MeSH *vocabulary* deals with every topic in medicine and its value in indexing is widely approved by the medical community. Two enhanced versions of this resource were actually incorporated into the database: the MeSH *tree-structure*, with the hierarchical organization, and the MeSH *annotated alphabetic list*, with related terms, synonyms, and different spellings of each descriptor. The MeSH

vocabulary comprises of 18,000 descriptors and 100,000 synonyms (Lowe & Barnett, 1994). This validated vocabulary was used in this project to perform two fundamental tasks: the MeSH terms list was compared with the terms list gathered from the filtering process for all the sub-collections and then the MeSH *tree-structure* was compared with the term structures generated by the classification algorithms. If variations of results can be shown in different sub-collections, then it is possible to assume that discourse aspects can certainly have a bearing on the application of filtering and classification algorithms.
- Stemming rules, (to be described later in this paper).

4.1. Experiment 1: Impact of discourse in the n-grams algorithm

The background selected in this study consisted of geology articles and a historical novel. These genres were selected because of the low degree of discourse overlaps with the document collection.

To improve the indexing results, a computer program checked the document to delete characters with low meaning in our system:

- Nonsense characters such as superscripts, numbers for footnotes and chapters, the "at" sign, emotions, etc.
- Punctuation signs (brackets, quotation marks, slashes, asterisks, etc).

A stemming algorithm (Díaz, Llorens, & Morato, 2002) was designed and implemented in our system to conflate all of the *n*-grams filtered terms into a normalized form. The system worked by finding the ending of each word, checking this ending against a list of term endings in the database, and replacing it with a normalized form: This process conflates all the words with the same stem into a single word. The information needed by the stemmers is included in two database tables: the "affixes" table and the "substitute-endings" table.

For instance, "virus" and "viruses" represent the same concept and can be normalized to "virus".

The terms obtained from the above process were checked against the descriptors of the controlled vocabulary. If one of these terms was found in the controlled vocabulary, the descriptor and its occurrences were updated in the database. If the term was neither a descriptor nor a stopword, it became a "candidate term". In order for this process to be successful, verbs were recognized and removed by comparing their normalized form against BNC and WordNet glossaries.

In this study, the value of the gram was finally chosen as five.

Once all the sub-collections were filtered, and the resulting terms were normalized, a total number of 1748 terms were selected by the computer to form the *n*-grams controlled vocabulary.

The indexing process of the sub-collections was done by simply referencing all the terms from any document found in the controlled vocabularies. The number of times every term occurred and its different positions within the documents were also stored in a database. The results showed that the indexing of the collection using the *n*-grams vocabulary produced references for 1748 terms, while the indexing using the MeSH vocabulary produced references for 3652 terms, including synonym referencing.

tf–idf was calculated for all the terms indexed (1748 and 3652), using the occurrence information calculated for every sub-collection. The formula applied was

$$\text{idf} = \log_2\left(\frac{N}{n_i} + 1\right)$$

Having N = total number of documents in the sub-collection and n_i = the number of documents in the sub-collection including the i term. This value was multiplied to the term frequency (tf) to get the final value. It is widely accepted that the higher the tf–idf is the better the term represents the information of the document. Therefore, comparing tf–idf for similar sub-collections, the first one indexed using *n*-grams vocabulary and the other one using MeSH, we could assess a certain degree of quality for *n*-grams. However, by comparing results for different sub-collections we could conclude how the algorithm is affected by discourse variables. In order to get a single tf–idf value for a whole sub-collection, the mean of all the terms' tf–idf was calculated. This measurement does not say a lot about terms but it should help to compare the behavior of the different sub-collections and vocabularies.

Table 5 shows the results obtained.

The following comments refer to the table of results:

1. The last column of the table (difference *n*-grams MeSH) shows that *n*-grams obtained better results in the four discourse variables studied. This is very interesting, because it leads us to say that instead of using very well known and established controlled vocabulary to index the documents of a particular domain, it seems to be better for retrieval purposes to automatically create a controlled vocabulary customized for the desired collection, and then index the collection using this vocabulary. The authors do not have a complete explanation for these results, although it could be related to the implicit weight system that *n*-grams performs when creating the vocabulary. *n*-grams has eliminated those "irrelevant" terms in every collection, thus, the application of tf–idf to these terms can imply better results.
   Looking at the results, with discourse variables we see that the highest difference was reached in the Register variable, (highest mean in the third column), which also has the best variance results.
   Regarding genre the differences, in absolute value, were highest in those genres that are less technical, and therefore more on the fringe of the medical domain ("Press" and "Popular-Science"). No clear explanations have been found for these results.
2. *Statistical mean calculations for every discourse variable*: The highest tf–idf values obtained by the *n*-grams vocabulary terms were found in "Press Articles" and "Notes" from genre, in "Press language" from register, "AIDS" and "Clinic Proteins" from domain terminology and "References" and "Abstract" from document structure. However, MeSH got better results in "Conference Proceedings" and "Popular-Science" from genre, in "Popular-Science language" from register, in "CJD" from Domain terminology and in "References" and "Abstract" from Document Structure.
   We must bear in mind that *n*-grams selects the accepted grams by gram frequency. This feature could explain why the algorithm obtained high values in different discourse variables. The highest value was obtained in "AIDS", and could be explained because of the impact that this disease had in the press and publishing in 1996 (year when the documents were published). The impact of terms like AIDS and HIV in the press was very high.

Table 5
Comparing tf–idf results for *n*-grams vocabulary and MeSH vocabulary

	n-grams average tf–idf	MeSH average tf–idf	Difference *n*-grams MeSH
Genre			
Press articles	2.19	0.98	1.21
Notes	2.19	1.43	0.76
Research articles	1.94	1.30	0.64
Conference proceedings	1.77	2.22	−0.45
Popular-science	2.02	2.69	−0.67
Genre mean	2.02	1.72	0.30
Genre variance	**0.03**	**0.40**	**0.53**
Register			
Scientific language	2.09	0.92	1.27
Press language	2.21	1.16	1.05
Popular-science language	1.93	1.2	0.73
Register mean	2.07	1.09	1.02
Register variance	**0.02**	**0.02**	**0.05**
Domain terminology			
AIDS	2.43	1.13	1.37
Clinic proteins	2.32	1.50	0.82
Hepatitis	1.86	1.10	0.76
CJD	1.79	1.87	−0.08
Domain terminology mean	2.10	1.40	0.72
Domain terminology variance	**0.08**	**0.10**	**0.27**
Document structure			
Discussion	2.12	1.05	1.07
Introduction	2.15	1.25	0.9
Results	2.06	1.2	0.86
Methods	2.03	1.7	0.33
References	2.23	2.20	0.03
Abstract	2.36	2.8	−0.44
Document structure mean	2.16	1.7	0.46
Document structure variance	**0.01**	**0.37**	**0.26**
All-variables mean	2.10	1.48	0.62
All-variables variance	0.03	0.32	0.37

Special mention must be given to some expected results. ''Abstracts'' and ''References'' (in the document structure discourse variable) obtained the highest results in both vocabularies, although MeSH values are more important. These results confirmed the common use in IR of these sections to reference documents. Regarding ''Abstracts'', the explanation matches the idea that ''Abstracts'' have the highest concentration of meaningful terms (high content-bearing words). This agrees with Loseeworks (1996). Losee suggested that some document sections, like ''Abstract'', contain more and better index terms than other sections. In other to explain ''References'' values, we must consider that *n*-grams and MeSH do not

usually focus on author's names, journals, etc. but on the title which can indeed be considered to be an abstract of the Abstract. Many databases used only title terms as retrieving elements.

We can see that the Document structure variable obtained almost the best values both in *n*-grams and MeSH vocabularies. It seems that structuring documents in a clear way is a very big help for retrieval purposes.

3. *Variability studies*: The most interesting results can be shown by studying variance values for *n*-grams. The results show that *n*-grams were almost unaffected by discourse variables when applied to a medical domain and compared through tf–idf results. Low figures can be seen in the intra-variables values of variance: e.g. no particular difference of *n*-grams results can be perceived when the algorithm was applied to register sub-collections or to any of the others. Indeed, the total variables variance value is also very low. However, MeSH did not react the same way, specially regarding genre and document structure. Register seems to be the only discourse variable that does not affect MeSH results, nor does it affect *n*-grams. What is more, the variance is exactly the same for *n*-grams and MeSH.

As a collateral result of this experiment, the authors thought that controlling the percentage of filtered terms gathered by *n*-grams and found in the MeSH vocabulary, could provide measures of the dependence of domain terminology on the other discourse variables: in particular, to study how close domains use different synonyms to describe the same concept. After filtering all the sub-collections the results showed that 498 out of the 1748 descriptors obtained through the *n*-grams process were found in the MeSH vocabulary, some 30%. The results are presented in Table 6.

Regarding domain terminology, we noticed the low correspondence (10%) of "Botanic proteins" terminology present in MeSH. However, when MeSH synonyms were taken into account, the "Botanic proteins" domain increased to 63%. The possible explanation for this behavior could be gathered from the percentage descriptors in the (D + S) column of Table 6. It seems that the "Botanic proteins" domain experts do not share the same descriptors as the medical experts. However, when including synonyms these results improve. The mean value for the descriptors found in the collection agreed with previous studies for particular domains (Bates, 1986).

The low number of terms gathered by "Press language" (in Register variable), helps to assess the conclusions of the experiment: "Press" uses only a short set of very often repeated medical terms.

Some results that could "a priori" be expected were confirmed. Regarding document structure results, terms extracted from the "Methodology Section" of the documents were usually found in MeSH under the generic heading "Investigative techniques". So were the more frequent terms in Abstract and Conclusion sections of the documents "Heterocyclic Compounds", "Investigative Techniques" and "Viruses". Many descriptors extracted from the "Reference" section were related to the geographical hypernyms; this is probably to do with the fact that all the domain terminologies represent pandemic diseases. This confirmation allows us to consider that future algorithms using only the MeSH descriptors under those headings will be able to automatically identify "Methodology" documents.

Table 6
Discourse influence in *n*-grams algorithm

	Total terms obtained with *n*-grams	Percentage of filtered terms matching MeSH		Difference: (D + S) − (D) (%)
		Comparing only with MeSH descriptors (D) (%)	Comparing with MeSH descriptors and synonyms (D + S) (%)	
Genre				
Press articles	579	20.6	70.3	49.7
Conference proceedings	330	13.9	51.8	37.9
Popular-science	434	14.3	50.5	36.2
Research articles	1546	18.4	50.3	31.9
Notes	680	16.2	45.6	29.4
Register				
Press language	382	15.7	52.9	37.2
Popular-science language	867	11.2	35.8	24.6
Scientific language	1515	13.9	32.0	18.1
Domain terminology				
Botanic proteins	70	10.0	62.9	52.9
CJD	239	13.0	52.7	39.7
Clinic proteins	565	13.8	43.0	29.2
Hepatitis	710	13.4	41.4	28
AIDS	1622	13.0	29.9	16.9
Document structure				
Discussion	526	30.8	86.5	55.7
Methods	447	14.3	44.7	30.4
Introduction	804	18.8	43.7	24.9
Abstract	387	17.1	40.1	23
References	523	15.3	32.3	17
Mean		15.76	48.13	

4.2. Experiment 2: Impact of discourse in the k-means algorithm

The structure of the experiment was as follows:

In order to get the number of clusters to be formed by *k*-means, the number of hierarchies in MeSH was calculated. 15 different root terms (headings) were found in MeSH. The next hierarchical level included 110 terms.

These 110 terms were considered the base of the 110 hierarchies from MeSH to be compared with the *k*-means clusters. In order to compare the MeSH hierarchy clusters with the *k*-means clusters, the following measures were defined:

- *a* value: Number of headings from *i*th hierarchy in MeSH present in *j*th *k*-means cluster,
- *b* value: Number of terms from *j*th *k*-means cluster NOT present in *i*th hierarchy in MeSH,

- *c* value: Number of headings from *i*th hierarchy in MeSH NOT present in *j*th *k*-means cluster.

In order to compare results, the resemblance coefficient of Jaccard (Romesburg, 1984) was calculated:

$$C_{ij} = \frac{a}{a+b+c}$$

This coefficient must be calculated for every combination of *k*-means/MeSH clusters. As *k*-means need to be provided beforehand with the number of clusters to be created, and MeSH has 110 clusters, the algorithm was run to create 4, 16, 32, 64 and 128 clusters.

In order to compare global results, the average of all the Jaccard coefficients was assigned as the representative for an entire sub-collection. The obtained results are shown in the next graphs. These are the results for four different linguistic variables, genre, register, domain terminology, and document structure.

The analysis of the previous graphs showed that *k*-means did not seem to be affected by genre (Fig. 4(a)) and Register (Fig. 4(b)) while it was affected by document structure (Fig. 4(c)) and Domain Terminology (Fig. 4(d)). The 4(c) graph shows that the "AIDS" sub-collection obtained a better behavior than the rest. We must bear in mind that the values gathered by AIDS in Table 5 were also higher. This could imply a correlation between good performance of *k*-means and tf–idf. This hypothesis could be supported by the results, comparing Table 5 and these graphs: "Press articles" for genre, "Abstracts" for document structures and "Press language" for register.

On the other hand, the descriptors from some domain terminologies, like "Hepatitis", clustered worse in comparison with MeSH than "AIDS" when the number of classes increased.

Regarding the impact of domain terminology in *k*-means, shown in Fig. 4(c), some conclusions could be given:

The difference comes from a worse behavior of "CJD" and "Hepatitis" rather than a better behavior of "AIDS" ("AIDS" values are more or less the same as the results of 4(a) and 4(b)). Possible reasons for these figures could be that (1) "CJD" contained fewer documents in the sub-collections and therefore *n*-grams performed worse, and (2) "Hepatitis" had a much more diffuse domain terminology than "AIDS", more centered (by 1996) in prophylaxis and epidemiology. (3) The highest number of "AIDS" descriptors gathered by *n*-grams (see Table 6) directly implied a better behavior of *k*-means for low numbers of clusters.

Fig. 4(d) showed that *k*-means is certainly affected by document structure discourse variable. As a general feature, the absolute values for Jaccard distance in this figure are lower than the rest of the values gathered for the other discourse variables (e.g. the highest value for "Abstract" is lower than 5.5). The explanation for these results must be linked to the idea that document structure variable implies that only parts of the documents are used as input for the algorithms.

Fig. 4(d) also points out that "Abstract" document structure got highest results when the descriptors formed a low number of clusters, although it had the worse results for high numbers of clusters. This strange behavior could be related to the following factors:

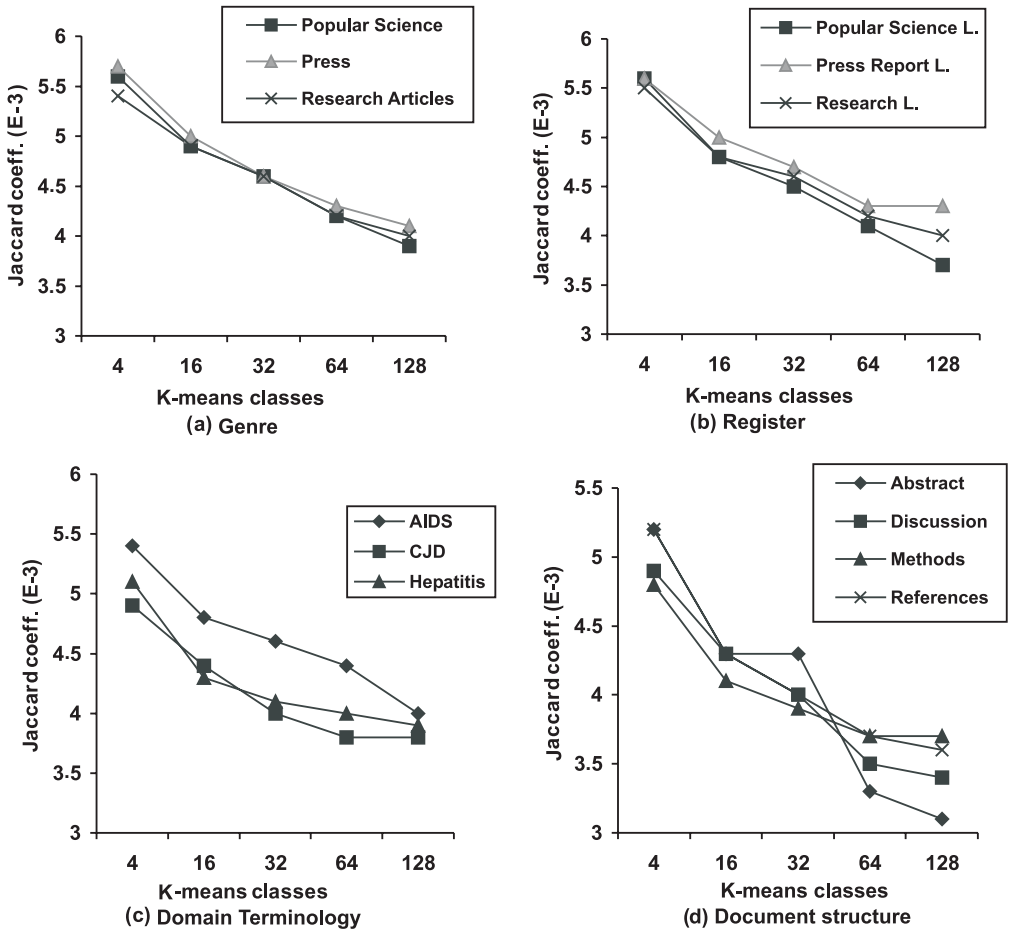


Fig. 4. *k*-means compared with MeSH. The horizontal axis indicates the number of classes selected in *k*-means. The graph shows the average of Jaccard resemblance coefficient for different discourse variables.

1. "Abstracts" contain very little ancillary information, that is, the descriptors are subject centralized. This property would affect *k*-means positively when the algorithm tries to create a low number of clusters, and negatively when the terms are forced to form many clusters.
2. "Abstract" usually has fewer descriptors than other document sections (see Table 6). The number of terms clearly affects the results of any classification algorithm.

We can conclude that once the optimal number of clusters is reached the results obtained are usually very bad if the algorithm over-classifies the terms.

4.3. Experiment 3: The impact of discourse in the Chen algorithm

The analysis using the algorithm developed by Chen coincided a lot with the previous method. Although, according to Díaz (Díaz et al., 1998), when this is applied to thesaurus construction the

difference between *k*-means and the Chen algorithm is that Chen detects the associations while *k*-means builds hierarchies of terms. The Chen algorithm tries to detect the degree of association between couples of words which appear associated in the documents.

The structure of the experiment was as follows:

The terms generated in *n*-grams were used as input into the Chen algorithm. The result was a group of binary associations between terms. These associations were checked against existing associations in MeSH.

We must also take into account that usual subject headings, like MeSH, are distinguished from the classical thesauri headings because the former has fewer associations among terms than the latter one. Therefore, associations from the Chen algorithm should present a poor coincidence with MeSH associations. This could explain the mediocre results shown in Table 7.

Table 7 shows the number of shared associations between the Chen algorithm and MeSH. It shows that higher results were obtained in "Research Articles" for genre, "Scientific Language" for register, "CJD" and "AIDS" for domain terminologies and "Discussion" for document structure.

Table 7
Shared items between Chen classification and MeSH

	Binary associations generated by Chen with terms existing in MeSH	Percentage of associations under the same heading in MeSH compared with the total of associations generated by Chen
Genre		
Research articles	496	11.3%
Conference proceedings	97	9.3%
Popular-science articles	143	7.7%
Notes	214	6.5%
News	110	3.6%
Variance		6.6
Register		
Scientific language	503	11.1%
Popular-science language	209	7.2%
Press language	140	3.6%
Variance		9.5
Domain-terminology		
CJD	20	15.0%
AIDS	503	11.1%
Clinic proteins	120	7.5%
Hepatitis	222	5.4%
Variance		13.4
Document-structure		
Discussion	459	15.0%
Methods	383	9.7%
Introduction	252	9.1%
Abstract	250	8.8%
References	156	5.8%
Variance		9.0

It can be observed that the results obtained by Chen had little in common with those obtained by *k*-means. Indeed, they seem to have an opposite behavior, at least for "News" in genre, "Press language" in register or "Abstract", "CJD" in domain terminology, and "References" in document structure.

The Chen algorithm indicates the probability that two terms could be associated in the same document. Therefore, the smaller the document is, the less probability we have of finding a significant association. This could explain the results obtained from small documents, like "News", "Abstracts" or "References".

Terms associations were also used to analyze the distribution of the information throughout the text. Each descriptor was referenced by its MeSH hypernym. Then, the occurrence of couples of hypernyms in the "Abstract", "Methodology", "Discussion", and "Reference" sections were evaluated (see Fig. 5).

Fig. 5 shows concrete results for the document structure discourse variable. Several rhombuses were placed between two heading associations. A rhombus can have different values: A for "Abstract", M for "Methodology", D for "Discussion", and R for "References" (Bibliography). When two headings have been found in a particular document section, a rhombus for this section will be placed between them.

Studying the results obtained, the following statements were presented in Table 8.

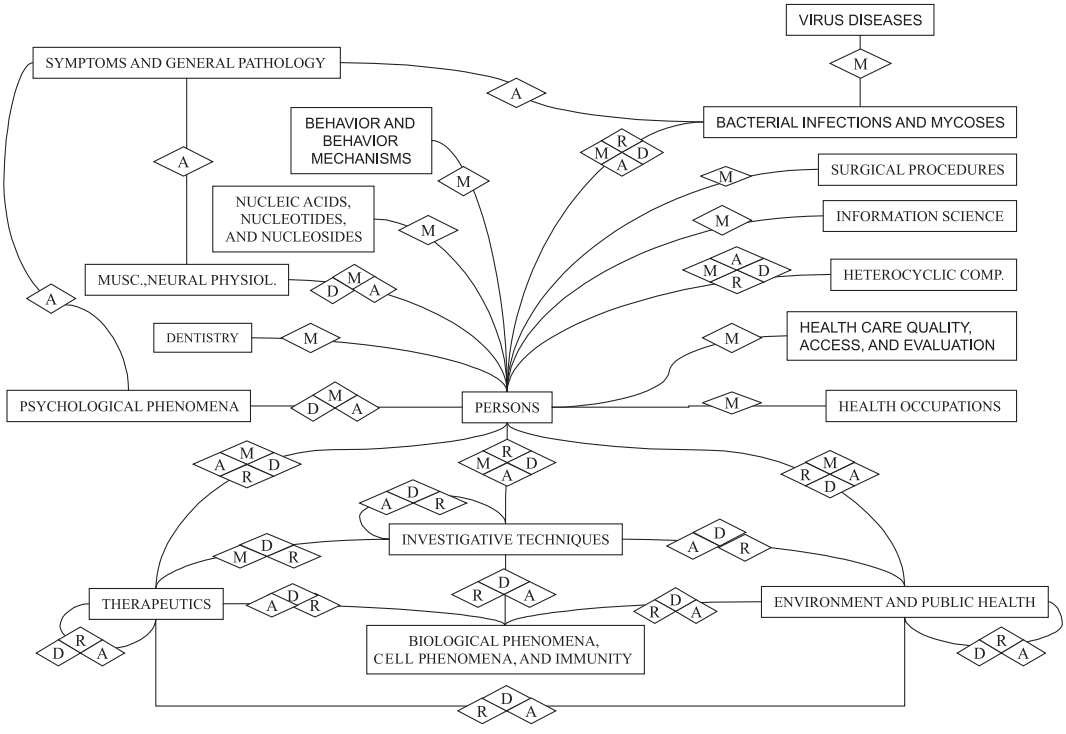


Fig. 5. The Chen algorithm: in our medical domain, we obtained the following relationships between pairs of terms in each document structure. A-abstract, M-methodology, D-discussion, R-references.

Table 8
Distribution of associations found by Chen's algorithm for different document sections

Groups of document sections	Percentage of associations found in all of the sections of the group (%)
"Abstract" + "References" + "Discussion"	29.6
"Methodology"	29.6
"Abstract" + "References" + "Discussion" + "Methodology"	18.5
"Abstract"	11.1
"Abstract" + "Discussion" + "Methodology"	7.4
"References" + "Discussion" + "Methodology"	3.7

The results presented in this table should be read in the following way: 29.6% of all the associations gathered by Chen's were found either in "Abstract", "References" and "Discussion" sections at the same time, while 11.1% were only found in the "Abstract" section.

We can see that "Discussion" and "References" often coincide (52%). This circumstance probably occurs due to rhetorical aspects of discourse (see Section 2.1). The argumentative language characteristic from the discourse section often used bibliographic references to support the assertions. On the other hand, "Methodology" and "Abstract" does not seem to appear together (26%) inside the lozenges.

This result demonstrates that the "Methodology" section is not correctly represented in the abstract section. For example, with "Virus diseases" the term is connected with "Bacterial Infections and Mycoses" by an M, we can conclude that both terms have been found in the Methodology section of documents.

5. Conclusions

The aim of this research project was to study the impact of genre, register, domain terminology and document structure in the Information classification and retrieval (ICAR) text algorithms. In order to achieve this objective, an experimental framework had to be defined where discourse dependant results could be calculated and contrasted.

IR systems usually only apply morphological and syntactical analyses without solving semantic ambiguity or coherence problems, and few studies have been made to measure the impact of discourse in IR systems.

ICAR algorithms give a different quality of results depending upon the discourse variables involved. Information science text analysis algorithms were initially expected to behave differently depending on context, though the results presented in this paper show that *n*-grams filtering algorithm does not seem to be affected by discourse variables. However, *k*-means and the Chen classification algorithms seem to be affected by them. This implies that those algorithms could certainly enhance their efficiency if context factors were taken into account.

It seems that a correlation can be found between tf–idf and *k*-means. High values of tf–idf usually imply better results for *k*-means.

This study also confirms the value of classical techniques and principles of Information Science that support the idea of giving more impact to abstract information in document indexing, or for references in bibliometrics (Callon et al., 1993). The lay out of document structures with a high

density of meaningful words is an essential value for indexing and in IR systems (Wormell, 1985). Some structures such as abstracts, conclusions, captions in figures and tables, beginnings of paragraphs, and titles, are mentioned in the text, because they provide relevant information (Losee, 1996), and these structures can therefore be used to extract valuable information (Wormell, 1985).

In this study we have compared patterns of language usage between an age old disease, hepatitis, and a very recent one, AIDS. This is probably why there is a greater number of different controlled terms in the hepatitis domain. The different behavior in *k*-means and *n*-grams of both diseases could also be due to this (see Fig. 5).

References

Amitay, E. (1998). Using common hypertext links to identify the best phrasal description of target web documents. *SIGIR'98 Post-Conference Workshop on Hypertext IR for the Web*, Melbourne, Australia.

Bates, M. (1986). Subject access in online catalogs: a design model. *JASIS, 11*, 357–376.

Beghtol, C. (2001). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science and Technology, 27*(2), 17–19.

British National Corpus (http://info.ox.ac.uk/bnc/, last check 1/11/01).

Bruce, N. J. (1983). Rhetorical constraints on information structure in medical research report writing. *ESP in the Arab world conference*. UK: Univ. Aston.

Callon, M., Courtial, J.-P., & Penan, H. (1993). *La scientométrie*. Paris: PUF.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translation to problematic networks: an introduction to co-word analysis. *Society Science Information, 22*, 191–235.

Chen, H., & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics, 22*, 885–902.

Cohen, J. (1995). Highlights: Language and domain-independent automatic indexing terms for abstracting. *JASIS, 46*(3), 162–174.

Díaz, I., Llorens, J., & Morato, J. (2002). An algorithm for term conflation based on tree structures. *JASIS, 53*(3), 199–208.

Díaz, I., Velasco, M., Llorens, J., & Martinez, V. (1998). Semi-automatic construction of thesaurus applying domain analysis techniques International. *Forum on Information and Documentation, 23*(2), 11–19.

Dijk, T. A. V. (1988). *News as Discourse*. Hillsdale, NJ: Erlbaum.

Egghe, L., & Roussau, R. (1990). *Introduction to informetrics: Quantitative methods in Library, Documentation and Information Science*. Amsterdam: Elsevier Science.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River: Prentice Hall.

Garfield, E. (1953). The relationship between mechanical indexing structural linguistics and information retrieval. *Journal of Information Science, 18*, 343–354 (1992, sent to the First Symposium on Machine Methods for Scientific Documentation (Johns Hopkins University, March 1953), the paper was rejected, but in 1992 the paper was published).

Gilyarevsky, R., Uzilevsky, G., & Moudrov, E. (1997). An automatic statistical classification of different types of journals. *International Forum on Information and Documentation, 22*(3), 24–35.

Haas, S. W., Sugarman, J., & Tibbo, H. (1996). A text filter for the automatic identification of empirical articles. *JASIS, 47*(2), 167–169.

Halliday, M. A. K. (1985). *Introduction to functional grammar*. London: Arnold.

Hearst, M., & Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the 16th ACM SIGIR conference on research and development in information retrieval*. NY: ACM.

Kando, N. (1997). Text-level structure of research papers: Implications for text-based information processing systems. Aberdeen: British Computer Society IR SG Annual Colloquium; 1997.

Karlgren, J. (1998). Stylistic experiments for information retrieval. In T. Strzalkowski (Ed.), *Natural language information retrieval*. Tomek: Kluwer.

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In: *Proceedings of COLING 94*, Kyoto.

Lavid, J. (1995). Towards a text type taxonomy: a functional framework for text analysis and generation. *Revista Procesamiento Lenguaje Natural, 16*, 29–43.

Lelu, C. (1993). *Modéles neuronaux pour l'analyse de données documentaires et textuelles*. Ph.D., Université de Paris, Paris.

Leydesdorff, L. (1997). Why words and cowords cannot map the development of the sciences. *JASIS, 48*(5), 418–427.

Llorens, J., Velasco, M., & Martínez Orga, V. (1997). Generación automática de representaciones de dominios. II Jornadas en Ingeniería de Software, JIS97, San Sebastián (Spain).

Llorens, J., Velasco, M., Morato, J., & Moreiro, J. A. (1998). Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros. *Revista de Procesamiento del Lenguaje Natural, 23*, 61–68.

de Looze, M. A., & LeMarié, J. (1997). Corpus relevance through Co-word analysis: an application to plant proteins. *Scientometrics, 39*(3), 267–280.

Losee, R. M. (1996). Text windows and phrases differing by discipline, location in document, and sysntactic structure. *Information Processing and Management, 32*(6), 747–767.

Lowe, H. J., & Barnett G. O. (1994). Understanding and using the medical subject headings vocabulary to perform literature searches. *JAMA*. 13, 271, No. 14, pp. 1103–1108.

Mitkov, R. (1998). The latest in anaphora resolution: going multilingual. *Revista Procesamiento Lenguaje Natural, 23*, 1–7.

Morato, J. (1999). *Análisis de las relaciones cienciométricas y lingüísticas en un entorno automatizado*. Ph.D. thesis. Universidad Carlos III de Madrid, Leganés, Madrid (Spain).

Neighbors, J. (1981). *Software construction using components*. Ph.D. thesis, Department of Information and Computer Science. Irvine: University California.

Nwogu, K. N. (1997). The medical research paper: structure and functions. *English Specific Purposes, 16*(2), 119–138.

Pêcheux, M. (1969). *Analyse automatique du discours*. Paris: Dunod.

Polanco, X., Grivel, L., & Royauté, J. (1995). How to do things with terms in informetrics: Terminological variation and stabilisation as science watch indicators. In *Proceedings fifth international conference on scientometrics and informetrics* (pp. 435–444). Learned Information, Medford (NJ).

Posterguillo, S. (1996). Is byte popular science? *Lenguas para fines especificos (V). Alcalá de Henares*: Publicaciones de la Universidad de Alcalá, pp. 425–432.

Prieto-Díaz, R. (1988). Domain analysis for reusability. In W. Tracz (Ed.), *Software reuse: emerging technology* (pp. 347–353). IEEE Computer Society Press.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. CA: Lifetime Learning Publications.

Schriffrin, D. (1994). *Approaches to discourse*. Oxford: Blackwell Publishers.

Skelton, J. (1994). Analysis of the structure of original research papers: An aid to writing original papers for publication. *British Journal General Practice, 44*, 455–459.

SMART project (ftp://ftp.cs.cornell.edu/pub/smart/, last check 1/11/01).

Spark Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Warner, A. (1994). The role of linguistic analysis in full-text retrieval. In *Challenges in indexing electr. text and images* (pp. 247–264). Medford: Learned Information.

Wormell, I. (1985). *Subject Access Project* (*SAP*). Improved Subject Retrieval for Monographic Publications. Ph.D. thesis, Lund: Lund University.