Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents

Elizabeth D. Liddy
Woojin Paik
Joseph K. Woelfel
School of Information Studies
Syracuse University
Syracuse, New York 13244-4100
liddy@mailbox.syr.edu
wpaik@mailbox.syr.edu
jkwoelfe@sunrise.acs.syr.edu

I. OVERVIEW

We are currently developing a system whose goal is to emulate a human classifier who peruses a large set of documents and sorts them into richly defined classes based solely on the subject content of the documents. To accomplish this task, our system tags each word in a document with the appropriate Subject Field Code (SFC) from a machine-readable dictionary. The within- document SFCs are then summed and normalized and each document is represented as a vector of the SFCs occurring in that document. These vectors are clustered using Ward's agglomerative clustering algorithm (Ward, 1963) to form classes in a document database. For retrieval, queries are likewise represented as SFC vectors and then matched to the prototype SFC vector of each cluster in the database. Clusters whose prototype SFC vectors exhibit a predetermined criterion of similarity to the query SFC vector are passed on to other system components for more computationally expensive representation and matching.

This classification system is part of DR-LINK, a larger document retrieval system being developed under the auspices of DARPA's TIPSTER Project (Liddy & Myaeng, 1991). DR-LINK will function as both a retrospective document retrieval system and a document dissemination system. It will be used in both modes to meet the information needs of a range of users and will handle millions of documents of various types. Our system is modular in design, consisting of a series of sub-systems, the first of which, the document classifier is presented in this paper. In brief, DR-LINK will first classify documents according to subject matter and then for those documents which are sufficiently similar in subject matter to a query, delineate the discourse- level organization of their content in order to focus the search for particular types of information on the appropriate sections of the document. Next we will use Relation-Revealing Formulae (Liddy & Paik, 1991), which rely on predictable linguistic features of text to detect a range of semantic relations (e.g. cause, purpose, location). The Relation-Revealing Formulae provide concept-relation-concept triples to the Conceptual Graph (CG) generator (Sowa, 1984) which will produce Conceptual Graph representations of document contents to be matched with Conceptual Graph representations of users' queries.

Although CG matching enable us to do fine-grained searching, it is computationally expensive. Such fine-grained representation is not necessary to determine, for instance, that a document on AIDS is not likely to be relevant to a query on terrorism. Therefore, we use our document classifier to produce a first rough cut of those documents which have the potential of matching a query as the

ISSN: 2324-9773

first of a two-stage model of retrieval. Because the document classifier is based on the implicit semantics of the words in the text, it offers an opportunity to successfully eliminate non-topic relevant documents during a preliminary stage without the attendant risks of clustering approaches which are based on non-semantic characteristics of documents.

II. DOCUMENT CLASSIFICATION PROBLEM

It would seem no longer necessary to make the case for automatic document classification given the known negative aspects of classification based on manually assigned keywords, namely imprecision, inefficiency, and inconsistency. There have, however, been a range of approaches to the task of automatic document classification with the most recent efforts applying the processing power of large parallel computers (Masand et al, 1992). However, approaches such as Masand et al (1992) require large manually coded training samples of up to 50,000 texts.

The task of automatic document classification remains a very difficult problem given the richness and variety of natural language. Imprecise classification of a document collection risks excluding from further consideration those documents which might match a query during a later, finer matching process. However, we believe that a subject-based document classifier which uses the intrinsic semantics of documents offers a means of partitioning a large heterogenous collection into smaller, more cohesive sub-collections, each of which constitutes a more homogenous collection on which to perform finer-grained matching procedures.

Although we agree in principle with Rasmussen (1992) that cluster analysis is not strictly the same as automatic classification because in clustering the classes formed are not defined a priori, but are determined by the entities assigned to them, we consider our clustering-based document partitioner to be an automatic classification system. As such, there are several vital factors that will determine its success as an automatic classification system: 1) the representation or attributes on which the entities will be classified; 2) the principle or theory by which these entities will be clustered, and; 3) the measure which will determine that the collection has reached the optimum level of clustering.

III. REPRESENTATION OF DOCUMENTS FOR CLASSIFICATION

We will cluster documents using the Subject Field Codes from *Longman's Dictionary of Contemporary English*. We use SFCs as an intermediate level representation of a text's contents. They are similar to a 'controlled vocabulary' representation, thereby taking care of the "synonymous phrasing" problem that plagues the use of natural language in information retrieval, but they do not require human assignment. We use SFC vectors to represent texts at a more abstract, conceptual level than the natural language text without the artificiality and expense of manual indexing with a controlled vocabulary.

A. Longman Dictionary of Contemporary English

The machine-readable dictionary which we are using in DR-LINK is *Longman's Dictionary of Contemporary English* (LDOCE), a British-produced learner's dictionary. LDOCE has been used

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

in a number of investigations into natural language processing applications (Boguraev & Briscoe, 1989) using the first edition (1978) of the dictionary. We are using the second edition (1987). We began working directly from the typesetters' tape, which we have cleaned up during research into the automatic extraction of semantic relations from dictionary definitions (Liddy & Paik, 1991) and have converted the data into a lexical database.

The 1987 edition of LDOCE contains 35,899 headwords and 53,838 senses, for an average of 1.499 senses per headword. The machine-readable tape of LDOCE contains several fields of information not visible in the hard-copy version, but which are extremely useful in natural language processing tasks. Some of these are relevant for syntactic processing of text, such as subcategorization codes, while others contain semantic information, such as the Box Codes, which indicate the class of entities to which a noun belongs (e.g. animate, abstract) or the semantic constraints for the arguments of a verb or an adjective, and the Subject Codes, which will be further detailed next.

B. Subject Codes in LDOCE

The Subject Codes are based on a classification scheme of 124 major fields and 250 sub-fields. According to Walker and Amsler (1986), the LDOCE subject fields are based on a classification scheme of Merriam-Webster, but no further description of the basis of the classification scheme is available. The fields and their sub-fields appear conceptually coherent, although they lack a consistent level of granularity across the major fields. For example, Education, Knots, Anatomy, and Cricket are all major fields. And although Sports itself is a major field, with five sub-fields, there are an additional fifteen major fields for various sports, such as Football, Golf, and Net Games, each with numerous sub-fields. This can pose a problem when the Subject Codes are used to represent text, and has provoked some researchers to impose additional levels on the basic two-level Subject Code hierarchy (Slator, 1991).

Subject Codes are manually assigned to words in LDOCE by the Longman lexicographers. There are two types of problems with the Subject Code assignments which become obvious when an attempt is made to use them computationally. First, a particular word may function as more than one part of speech and each word may also have more than one sense, and each of these entries and/or senses may be assigned different Subject Codes. The entries for 'acid' in Fig.1 are taken from the LDOCE tape and demonstrate a fairly simple example of this problem.

HEADWORD	PART-OF-SPEECH	SUBJECT FIELDS
acid	noun	SIzc [Science, chemistry] DG [Drugs (not pharmaceutical)]
acid	adjective	FOzc [Food, cookery] XX [General]

Fig 1: LDOCE entry with Multiple Parts of Speech and SFCs

ISSN: 2324-9773

If an NLP system cannot ascertain either the grammatical function or sense of a word in the text being processed, all Subject Codes for all entries for an orthographic form must be considered. However, our system incorporates automatic means for choosing amongst the LDOCE syntactic categories and choosing amongst the senses, thereby limiting which Subject Codes are assigned to each word in a given text.

There is also the possibility that no Subject Code has been assigned to a word or any of its individual senses. Of the 53,838 senses in LDOCE '87, 51,383 or 95% have Subject Codes. Of these, however, 27,273 are coded XX for the General class and therefore provide no useful semantic information. The absence of Subject Codes or the presence of only the General class code poses a problem when word-by-word disambiguation is desired, but when the task is to arrive at a summary semantic representation of the text, the law of large numbers appears to take over. For although only 24,110 senses (45%) in LDOCE have the more informative, domain-specific codes, this appears to be sufficient for the texts we have processed to make the task of text classification quite reasonable. In the future, we will investigate ways in which we can use sentence context and the correlation matrix to suggest appropriate SFCs for those words that do not have SFCs in LDOCE. However, the cases in which a word's senses have different SFCs impact more immediately on our attempts at classification, since the most frequently used words in our language tend to have many senses and therefore, multiple Subject Codes.

C. Other Work Using Subject Codes

Walker and Amsler, who were the first to make use of the domain information represented by Subject Codes, have reported on a somewhat similar attempt to utilize the Subject Codes to determine the subject domains for a set of texts (1986). However, they used the most frequent Subject Code to characterize a document's content, whereas we represent a document by a vector of frequencies of Subject Codes for words in that text. We find that our research efforts strongly support the suggestions made by Walker and Amsler concerning ways to refine the representation of text using Subject Codes.

Slator (1991) has taken the original 124 Subject Codes and added an additional layer of seven pragmatic classes to the original two-level hierarchy. These are communication, economics, entertainment, household, politics, science and transportation. He has found the reconstructed hierarchy useful when attempting to disambiguate multiple senses and Subject Codes attached to words. His metric for preferring one sense over another relies on text-specific values, whereas we add corpus correlation values as a further stage in the disambiguation process.

Krovetz (1991) is exploring the effect of combining the evidence from Subject Codes with evidence from morphology, part of speech, subcategorization and semantic restrictions for selection of the correct sense. His goal is to represent documents by the appropriate senses rather than just the orthographic forms of words, for use in an information retrieval system, .

D. Use of Subject Field Codes for Classification

The Subject Codes are a sub-field of the Definition Code field (DFC) on the LDOCE tape which we have extracted, relabeled as Subject Field Codes (SFCs) and made a separate field in the

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

LDOCE lexical database developed at Syracuse University. In our system, each sentence, paragraph, and document can be automatically represented as a vector of the normalized frequencies of the SFCs for that unit's constituent words. This sounds like a very simple and reasonable procedure and in fact would be, if it were not that many of the words in LDOCE have multiple grammatical categories and multiple senses and therefore multiple SFCs.

IV. CLUSTERING

Of the many clustering algorithms available, we have chosen Ward's minimum-variance method to use for automatic document classification on the basis of each document's SFC vector. We first focused on selecting an agglomerative clustering technique which "forms clusters by progressive fusion" (Michalski & Stepp, 1990) instead of either a divisive technique or a direct technique. After considering the number of documents in our database, a divisive technique was not deemed realistic because its computational cost is so high. The direct techniques require that the optimum number of clusters be established a priori, and we do not find this an intellectually reasonable parameter. After testing several agglomerative clustering methods, such as average linkage and centroid methods, we found Ward's minimum-variance method provided the most satisfactory performance with our sample data.

We chose Ward's because it doesn't exhibit a bias towards "chaining" like the single linkage method. Its bias is in the opposite direction. Ward's tends to form roughly spherical clusters of approximately equal numbers of members (SAS, 1988). Thus, the documents within a cluster should have a high degree of similarity. Also, Ward's does not tend to cluster together "non-conformist" documents that are not similar but which are clustered together simply because they don't belong to other clusters such as the case is with the average linkage method.

V. SYSTEM DESCRIPTION

In order to successfully convey a sense of exactly how the document classifier functions, the following detailed procedural description is provided:

A. Subject Field Coding of Documents

For each of the documents, the following stages of processing are done to generate vector representations of each document:

In **Stage 1** processing, we run the documents and query through POST, a probabilistic part of speech tagger (Meeter et al, 1991) loaned to us by BBN where it was developed. We use the tri-tag model of POST which "predicts the relative likelihood of a particular tag given the two preceding tags" (p. 961). We do not use POST for the purpose of producing a parse of the sentence, but rather to enable us to limit the SFCs of a word to those of the appropriate syntactic category of each word as determined by POST. Since many words in a general corpus such as WSJ can function as more than one part of speech, the inclusion of POST has reduced the number of SFCs

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop. 83-100. doi:10.7152/acro.v3i1.12598

that need to be further considered for sense disambiguation by an average of 60% in comparison with the case when the part of speech of each word was not resolved.

Stage 2 processing consists of retrieving SFCs of each word's correct part of speech from the lexical database. The SFC retrieval process utilizes a modified version of WordNet's exception dictionary (Beckwith & Miller, 1990) and the Kelly & Stone (1975) stemming algorithm. As a special case, if a word is a hyphenated word and no entry has been found with a POST assigned part of speech in the lexical database, the system removes the hyphen and searches the conjoined result in the lexical database. If not found, the system separates the words and assigns part of speech to each composite part using POST, and then these two words are looked up in the lexical database.

At **Stage 3** we begin sense disambiguation, using local sentence-level context-heuristics. In this research, we equate sense disambiguation to automatic determination of a single word's correct SFC. We begin with context-heuristics because empirical results have shown that local context is used successfully by humans for sense disambiguation (Choueka & Lusignan, 1985) and context-heuristics have been experimentally tested in Walker & Amsler's (1986) and Slator's work (1991) with promising results. The input to Stage 3 is a word, its part-of- speech tag, and the SFCs of each sense of that grammatical category. For some words, no disambiguation may be necessary at this stage because the SFCs for the part-of-speech of the input word may all be GENERAL or there may be no SFCs provided by LDOCE. However, for the majority of words in each sentence there are multiple SFCs, so the input would be as seen in Figure 2.

State	n	POLITICAL SCIENCE[4], ORDERS
companies	n	BUSINESS, MUSIC, THEATER
employ	v	LABOR, BUSINESS
about	adv	-
one	adj	-
billion	adj	NUMBERS
people.	n	SOCIOLOGY, POLITICAL SCIENCE[2], ANTHROPOLOGY

Fig 2: Subject Field Codes & Frequencies (in Superscript) for Words as one part-of-speech

To select a single SFC for each word in a sentence, Stage 3 uses an ordered set of heuristics. First, the SFCs attached to all words in a sentence are evaluated to determine at the sentence level: 1) whether any words have only one SFC assigned to all senses of the word; 2) the SFC which is most frequently assigned across all words in the sentence. Each sentence may have more than one unique SFC as there may be more than one word whose senses have all been assigned a single SFC. In Figure 2, NUMBERS is a unique SFC, being the only SFC assigned to the word 'billion' and POLITICAL SCIENCE is the most frequently assigned SFC for this sentence. We think that the unique SFCs and the most frequently occurring SFC are good local determinants of the subject domain of the sentence. We have established the criterion that if no SFC has a frequency equal to or greater than three, we do not select a frequency-based SFC for that particular sentence. Our preliminary test results show that SFCs with a within-sentence frequency less than three do not accurately represent the domain of the sentence.

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

The second step in Stage 3 evaluates the remaining words in the sentence and chooses a single SFC for each word based on the locally-important SFCs determined in step one of this stage. The system scans the SFCs of each remaining word to determine whether the SFCs which have been identified as unique or most frequent occur amongst the multiple SFCs assigned to each word by LDOCE. In Figure 2, for example, POLITICAL SCIENCE would be selected as the appropriate SFC for 'people' because POLITICAL SCIENCE was determined in step 1 to be the most frequent SFC value for the sentence.

Stage 4 incorporates two global knowledge sources to complete the sense disambiguation task begun in Stage 3. The primary source is a 122 x 122 correlation matrix computed from the SFC frequencies of the 442,059 words that occurred in a sample of 977 WSJ articles and so reflects stable estimates of SFCs which co-occur within documents. The second source is the order in which the senses of a word are listed in LDOCE. Ordering of senses in LDOCE is determined by Longman's lexicographers based on frequency of use in the English language.

The correlation matrix was computed by SAS using SFC output of the 977 WSJ articles from Stage 2 (each document is represented by a vector of SFCs of the senses of the correct part-of- speech of each word as determined by POST). The observation unit is the document and the variables being correlated are the 122 SFCs. The scores for the variables are the within- document frequencies of each SFC. There are 255,709 scores across the 977 articles on which the matrix is computed. The resulting values in the 122 x 122 matrix are the Pearson product moment correlation coefficients between SFCs and range from a +1 to a -1, with 0 indicating no relationship between the SFCs.

The output matrix is consulted during Stage 4 processing to determine the correlation coefficient between two SFCs and serves as the more global, document-level data on which we attempt to select one SFC for each word not disambiguated in Stage 3. The correlation coefficients are quite intuitively reasonable, as can be seen in Figure 3 where the ten highest correlations are listed. The unexpected correlation between LAW and BUILDING is due to the highly frequent usage of the word 'court' which has SFCs for both LAW and BUILDING and contributes greatly to the high correlation between the two SFCs.

Co-efficient	SFC-1	SFC-2
.91314	NET GAMES	COURT GAMES
.80801	ECONOMICS	BUSINESS
.73958	SOCIOLOGY	LAW
.73654	THEATER	ENTERTAINMENT
.72199	THEATER	MUSIC
.71428	PLANT NAMES	AGRICULTURE
.70844	ANIMAL HUSBANDRY	AGRICULTURE
.70271	AGRICULTURE	BUSINESS
.68600	LAW	BUILDING
.68177	GAMBLING	CARD GAMES

Fig. 3: Highest Correlations Between SFCs Based on 255,709 SFC Frequencies

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

In Stage 4, one ambiguous word at a time is resolved, accessing the matrix via the unique and most frequent SFCs determined for a sentence in Stage 3. The system evaluates the correlation coefficients between the unique/most frequent SFCs of the sentence and the multiple SFCs assigned to a word to determine which of the multiple SFCs has the highest correlation with the unique and/or most frequent SFCs. The system then selects that SFC as the unambiguous representation of the sense of the word.

We have developed heuristics for three cases for selecting a single SFC for a word using the correlation matrix. The three cases function better than handling all instances as a single case because of the special treatment needed for words with the less-substantive GENERAL (XX) or CLOSED SYSTEM PART OF SPEECH (CS) codes. For the two cases where there are XX or CS amongst the SFCs, we take order of the SFCs into consideration, reflecting the fact that the first SFC listed is more likely to be correct, since the most widely used sense is listed first in LDOCE. So, to overcome this likelihood, a more substantive SFC listed later in the entry must have a much higher correlation with the sentence-determined SFC.

In an attempt to clarify the description, we will refer to a word's multiple SFCs that the system must select amongst as word-attached SFCs and the unique and most-frequent SFCs that were established at the sentence level in Stage 3 as sentence-determined SFCs.

Case 1- Words with **no** XX or CS SFCs:

If any word-attached SFC has a correlation greater than .6 with any one of the sentence-determined SFCs, select that word-attached SFC.

If no word-attached SFC has such a correlation, average the word-attached SFC and sentence-determined SFCs correlations, and select the word-attached SFC with the highest average correlation.

Case 2 - Words with XX or CS listed first:

Select the XX or CS unless a more substantive SFC further down the list of senses has a correlation with the sentence-determined SFCs greater than 0.6.

Case 3 - Words where XX or CS is not the first listed SFC:

Choose the more substantive SFC if it has a correlation greater than 0.4.

Figure 4 presents a sample sentence which illustrates how the heuristics use the correlation matrix values to select correct SFCs for examples of Cases 1,2 and 3. For this sentence, BEAUTY CULTURE, CALENDAR, and ECONOMICS were selected at Stage 3 as unique SFCs, based on being the sole SFC assigned to 'cosmetics', 'November', and 'financing' respectively. Therefore, when a SFC needs to be selected for 'giant', Case 3 says that the more substantive SFC (here LITERATURE) must have a correlation greater than .4 with a unique SFC to be selected, therefore the GENERAL SFC will be chosen. This is the correct choice.

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

WORD	POS	SFC	CORRELATION WITH UNIQUE SFCs
He	pro		
acquired	v	GENERAL	
the	det		
cosmetics	n	BEAUTY CULTURE	
giant	n	LITERATURE, GENERAL	LITERATURE-BEAUTY CULTURE: .328 LITERATURE-CALENDAR: .334 LITERATURE-ECONOMICS: .162 -> Select GENERAL by Case 3
in	prep		
November,	n	CALENDAR	
financing	v	ECONOMICS	
in	prep		
the	det		
transaction	n	GENERAL	
with	prep		
junk	n	GENERAL, DRUGS, NAUTICAL	DRUGS-BEAUTY CULTURE: .317 DRUGS-CALENDAR: .329 DRUGS- ECONOMICS: .269 NAUTICAL-BEAUTY CULTURE: .3.76 NAUTICAL-CALENDAR: .434 NAUTICAL-ECONOMICS: .378 -> Select GENERAL by Case 2
bonds	n	GENERAL	
floated	v	GENERAL, BUSINESS	BUSINESS-BEAUTY CULTURE: .376 BUSINESS-CALENDAR: .545 BUSINESS-ECONOMICS: .808 -> Select BUSINESS by Case 2
by	prep		
DBL, Inc.	prop		

Fig. 4: Sample sentence exemplifying correlation-matrix heuristics

In the case of 'junk', since GENERAL is listed first, either DRUGS or NAUTICAL must have a correlation coefficient of at least .6 to be selected over GENERAL. Since neither do, the correct choice of GENERAL is made. For the case of 'floated', the same logic applies, but since BUSINESS has a correlation of .808, BUSINESS is selected over the first occurring GENERAL.

Our SFC disambiguation procedures were tested on a sample of 1638 words from WSJ which had SFCs in LDOCE. The system implementation of the disambiguation procedures was run and a single SFC was selected for each word. These SFCs were compared to the sense-selections made by an independent judge who was instructed to read the sentences and the definitions of the senses of each word and then to select that sense of the word which was most correct. Figure 5 summarizes the overall results (att. = attempts, cor. = correct) presented according to the main source of knowledge used in the disambiguation process (Full details of the disambiguation process and testing are available in Liddy & Paik, 1992).

ISSN: 2324-9773

Local Heuristics			Domain Correlations			Frequency Ordering			Total		
att.	cor.	%	att.	cor.	%	att.	cor.	%	att.	cor.	%
1134	1032	91	268	206	77	236	219	93	1638	1457	89

Fig. 5: SFC Disambiguation Results Using Multiple Sources of Knowledge on 1638 Words

Stage 5 processing produces a representation consisting of a vector of SFCs and their frequencies for each document and for the query. At this point the two non-substantive SFCs (CLOSED SYSTEM PART OF SPEECH and GENERAL) are removed from the SFC vector sums, since these contribute nothing to a text's subject content representation.

In **Stage 6**, the vectors of each document and the query are normalized using Sager & Lockeman (1976) term weighting formula in order to control for the effect of document length. The resulting, normalized document vectors are passed to the next clustering process. Formulae are provided in Figure 6.

Sager (1976) term weighting $= \frac{f_{in}}{K_n}$

where f_{in} = frequency of SFC i in document n

k_n = number of tokens (SFC occurrence) in document n $\left(= \sum_{i=1}^{M} f_{in} \right)$

M = number of SFC (ie. 122)

Fig. 6: Formula for Term Weighting Scheme

B. Clustering of Documents

All agglomerative clustering algorithms, such as Ward's, use essentially the same algorithm (Zupan, 1982). First, a distance matrix is calculated between each of the documents. Thus, if we have n documents, we make an n x n matrix that contains the distances between every possible combination of documents.

At the start of the clustering process, each of the documents can be thought of as being the only member of a single cluster. Thus, if you start with n documents you also start with n clusters. The two clusters that are most similar according to the distance measure are then joined to form a single cluster. Clusters are then represented by the average of all the vector representations of the documents they contain.

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

Next, a new distance matrix is calculated and again the two closest clusters are joined. This process repeats until all documents are merged into a single cluster.

Ward's method doesn't use a Euclidean distance measure to form the matrix. Instead of joining the two clusters that are closest in a Euclidean sense, Ward's method joins the two clusters that cause what Ward refers to as the least "loss of information". Ward measures lost information as the error sum of squares (ESS) given by the formula in Figure 7.

$$D_{KL} = B_{KL} = \| \bar{X}_K - \bar{X}_L \|^2 / (1/N_K + 1/N_L)$$

where D_{KL} = dissimilarity measure between cluster K and L

$\bar{X}_K$ = mean SFC vector for cluster K

$\bar{X}_L$ = mean SFC vector for cluster L

N_K = number of documents in cluster K

N_L = number of documents in cluster L

Fig. 7: Ward's Minimum Variance Method (SAS, 1988)

The above distance formula is equal to the sum of squares between the two clusters.

For every cluster an ESS can be calculated. These are then added to produce a total ESS across all groups. The objective of Ward's clustering is to minimize this total ESS. Thus, instead of joining clusters that are closest in terms of a Euclidean distance measure, Ward's method joins the clusters that cause the smallest increase in the total ESS. Another way of thinking of this is that Ward's method attempts to minimize the variance within clusters.

Hierarchical clustering approaches do not predetermine the number of clusters and there is currently no agreed on formula for determining the "correct" number of clusters (SAS, 1988). Criteria can be set for hierarchical clustering systems such that every possible number of clusters is created, from one-per-document down to a single cluster for the whole set of documents. Each clustering system, therefore, needs to determine the optimal number of clusters based on the purpose of the system, the type of documents being clustered, and the clustering algorithm being used. We established a criterion by examining the results of a sample-run and evaluating the content of each of the clusters as they were joined together. We started with the first two documents joined and looked to see if the documents joined were similar enough to form a cluster. We continued this process as more and more clusters were joined until the system reached a point where dissimilar clusters were being joined. We chose the value just prior to this point as the minimum number of clusters that adequately represent the data. In our view, when forty-eight

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

clusters had been created, the system had reached an optimum level of clustering. At that point, 75% of the variance was explained. 75% may not be the absolute value for all systems, but for our sample database it was the level at which the internal coherence of our clusters began to disintegrate. More research is necessary to determine if there are simple formulae or rules of thumb for estimating a reasonable number of clusters for other collections or whether the 75% variance explained holds across collections.

VI. TESTING OF AUTOMATIC CLASSIFICATION SYSTEM

To determine whether the SFCs could serve as a useful representation on which clustering of a large set of documents could be based, and whether Ward's approach was appropriate, a sample collection was created from the *Wall Street Journal* (Dow Jones, 1989). The collection contains a total of 250 documents. Fifty of these documents had been judged relevant to a query on the topic of business mergers and acquisitions during an earlier experiment by two out of three judges who read and evaluated 300 WSJ articles. To create a more heterogenous collection, an additional 200 documents were selected from the same WSJ database by viewing just the headlines and choosing those which appeared to cover a broader range of topics. The resulting sample of 250 documents and the one test query were processed according to the procedures described above.

Using this sample set, two types of evaluation of the results were conducted.

First, we evaluated our automatic classification scheme qualitatively. By observing what documents our system grouped together at each level in the hierarchy, we could see if these groupings made intuitive sense. Would a human classifier agree with these groupings? Do documents get classified into one cluster when they are actually more coherent with the documents in another cluster? What are the characteristics of the groupings developed using our method? The goal of this type of qualitative evaluation was to simply inform us whether SFC representations had the potential we had envisioned for classifying documents.

A second test of our automated classification was its performance in a retrieval task. To evaluate this capability, we compared retrieval results using, first, a similarity measure between the query and each of the individual documents to compute rankings of which documents were most similar (closest) to the query. Next we computed a similarity measure between the query and each of the clusters. We accomplished this by averaging the vectors of the documents comprising each of the 48 clusters to produce a prototype vector of each cluster (the centroid of each cluster). We then calculated the similarity between each of the prototypes and the query and ranked the clusters based on similarity to the query.

VII. RESULTS

A. Qualitative Evaluation

A qualitative analysis of the clusters revealed that the use of SFCs combined with Ward's clustering algorithm resulted in meaningful groupings of documents that were similar across concepts not directly encoded in SFCs. Two examples: all of the documents about AIDS were classified

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

together. Secondly, all of the documents about the hostages in Iran were classified together even though proper nouns are not included in LDOCE and the word 'hostage' is tagged with the same SFC as hundreds of other terms. What appears to happen with the SFC representation of documents is that relatively equal distributions of words from the same sets of SFCs are found in documents about the same or similar topics.

By examining two clusters in detail, some fine points of the results can be pointed out. The first cluster, Cluster 17 (Figure 8) can be generally characterized as consisting of documents about airlines. The leaf nodes are tagged with the headlines from the WSJ articles they represent.

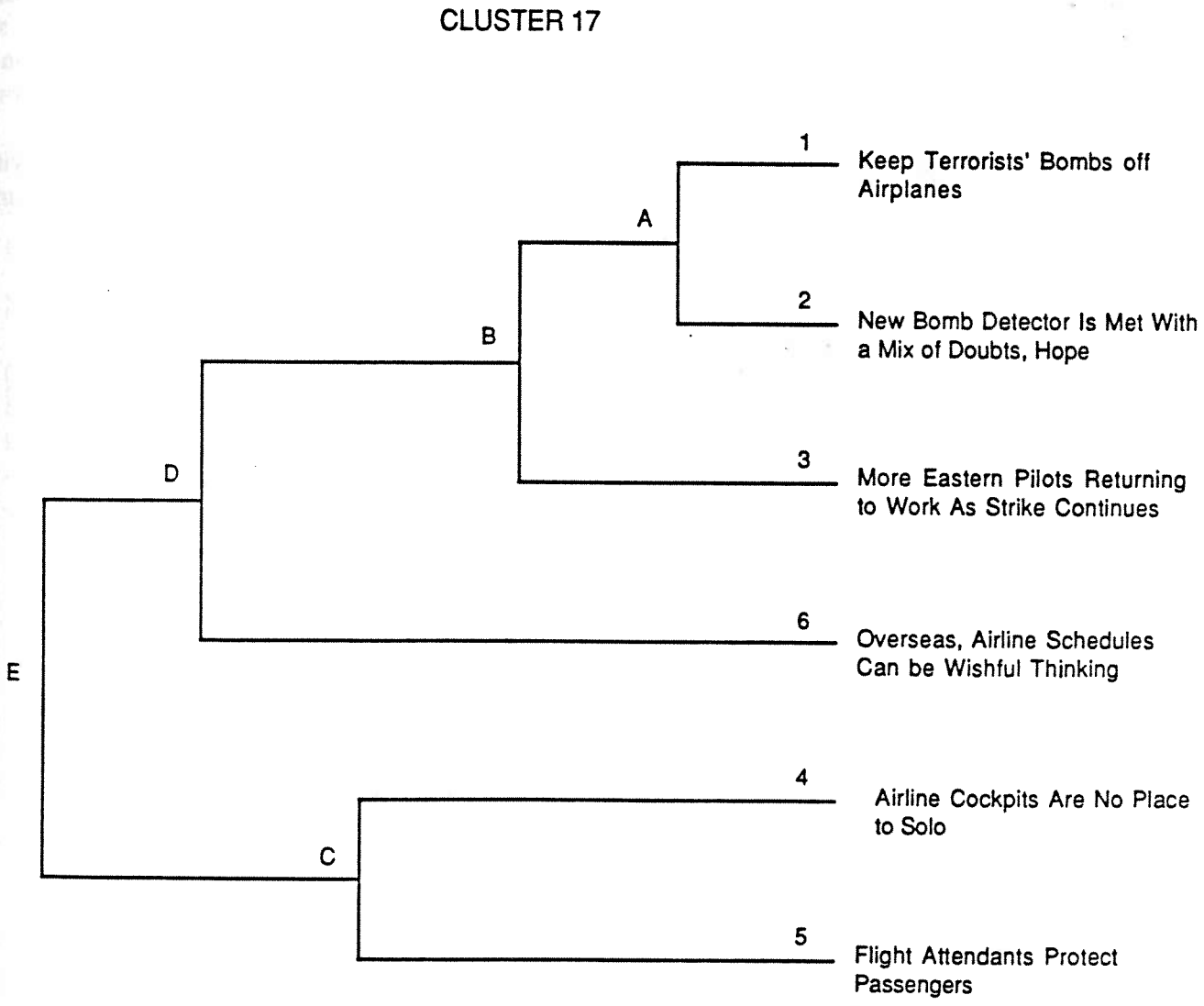

Figure 8

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

The two most closely associated documents (1 and 2) form sub-cluster A. These two documents are concerned with measures being taken to keep bombs and terrorists off airplanes. At the other end of cluster 17, documents 4 and 5 form sub-cluster C. Both of these documents are concerned with the training of airline employees (pilots and flight attendants) which prepares them to deal with airborne emergencies. The most anomalous document in Cluster 17 is document 3, which is concerned with the pilots' strike at Eastern Airlines. What appears to cause this somewhat odd grouping within the larger, consistent grouping of documents, is that of course it is about airlines and the word 'strike' has the same SFC as words such as 'security', 'bombs', etc. Document 6's headline does not clearly convey that the article is actually about delays in airline schedules caused by the need for new security measures in foreign airports. So its combining with the Clusters A and B is consistent based on document content.

The second cluster, Cluster 32 (Figure 9) contains eleven documents all concerned with the various aspects of medical treatment. Documents 1 and 2 report results of new studies of the effects of the drug, AZT on AIDS. Documents 3 and 4 are also about tests of drugs for AIDS and Parkinson's Disease. They comprise sub-cluster B and combine with the very similar sub- cluster A to form new sub-cluster C. Documents 7, 9, 10, and 11 are also about drug testing, but focus slightly more on the activities of groups trying to get access to a new drug or the activities of the companies which produce the drugs. At the other end of Cluster 32, are three documents concerned more with the economic side of medical care by hospitals which form sub- cluster H and eventually combine with the drug-research articles, but at a higher point in the hierarchy.

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

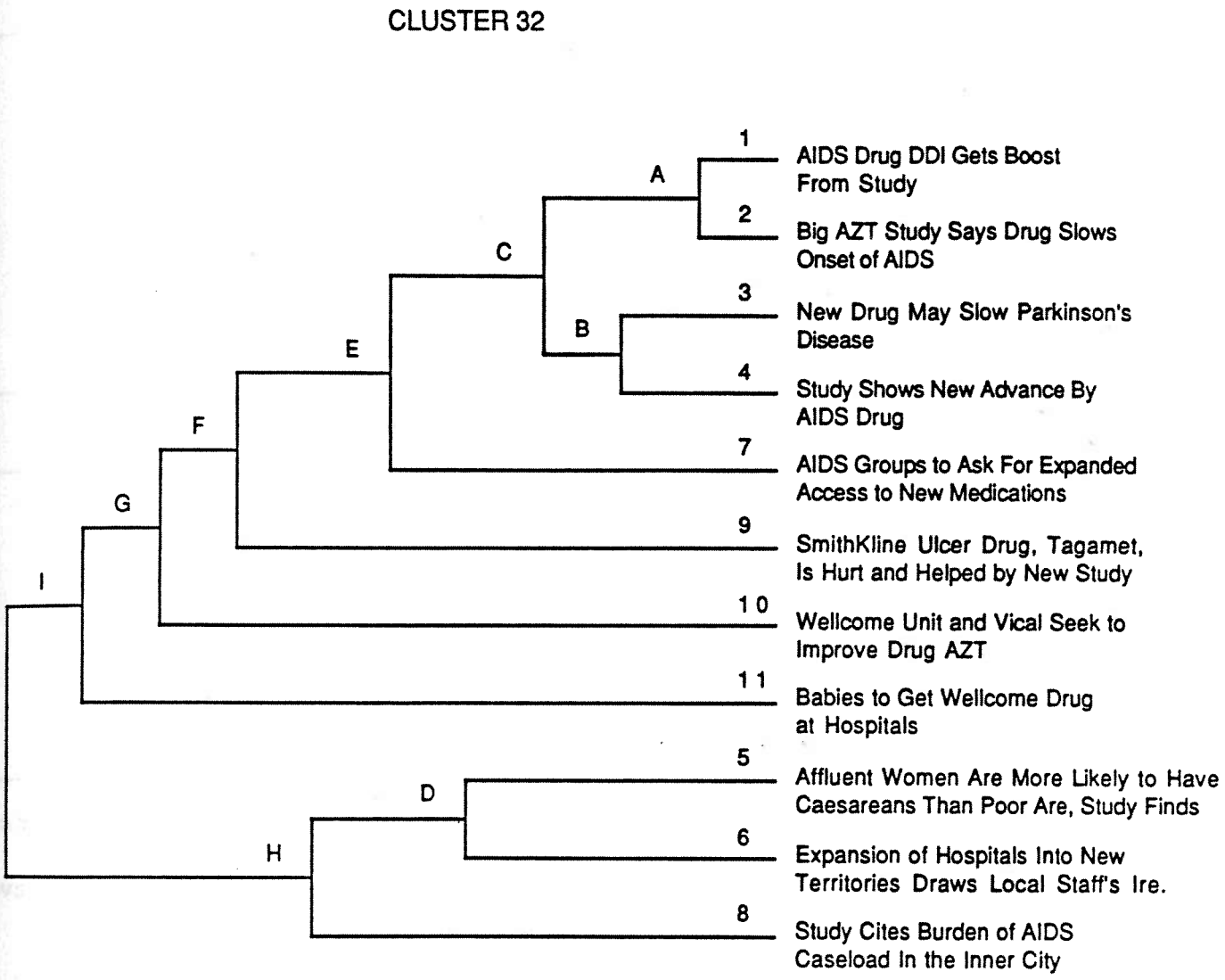

Figure 9

In summary, this intellectual interpretation of Clusters 17 and 32 strongly supports the results of the automatic clustering algorithm. Many more of these individual clusters could be detailed, but the point to be made is that the clusters cohere at all levels, and even at the criterion we have set, there are few documents which do not seem to fit naturally in the cluster to which they are assigned by the system.

B. Retrieval Experiment

Based on the collection of 250 documents from the WSJ, 48 clusters were created using Ward's minimum variance method. Among these 48 clusters, 17 clusters contained at least one document

ISSN: 2324-9773

relevant to the test query. Clusters were ranked by measuring similarity between the query and the centroid of each cluster using Sager & Lockemann's (1976) method (Figure 10).

$$S_{xy} = \frac{\sum X_i Y_i}{\sum X_i^2 + \sum Y_i^2 - \sum X_i Y_i}$$

X_i = weight of SFC i in the cluster (X)

Y_i = we ight of SFC i in the query (Y)

Fig. 10: Sager & Lockemann (1976) Similarity (S_{xy}) Measure

Figure 11 presents precision values at 4 recall points using the cluster as the unit of analysis.

Recall	Precision
at 0.25	1.00
at 0.50	1.00
at 0.75	0.94
at 1.00	0.65

Fig. 11: Recall-Precision based on 48 clusters and one test query

To evaluate how clustering of documents performed in the context of information retrieval, we can compare the recall-precision results based on clusters as shown in Figure 11 with the recall-precision result based on traditional matching of individual documents to the test query. However, the direct comparison between cluster-based recall-precision and traditional document-based recall-precision does not seem to be accurate enough as these results lack a common unit of analysis. Therefore, we computed new recall-precision values for the cluster- based information retrieval situation by assigning every document within a cluster the same rank (e. g. all 12 documents in the cluster which was ranked 7th in similarity to the test query receive the rank of 7). This method of comparison enables us to transform the ranked list of clusters into a ranked list of documents and thus make the document the unit of analysis for both situations. Figure 12 shows precision values at 4 recall points, based on the above cluster-to-document transformation.

Recall	Precision
at 0.25	1.00
at 1.50	1.00
at 0.75	0.80
at 1.00	0.42

Fig. 12: Recall-Precision based on 250 documents assigned tied-ranks within clusters

ISSN: 2324-9773

Liddy, E., Paik, W., & Woelfel, J. (1992). Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents. 3rd ASIS SIG/CR Classification Research Workshop, 83-100. doi:10.7152/acro.v3i1.12598

Figure 13 shows precision values at 4 recall points for the results from the traditional, unclustered, document-based information retrieval situation. The recall-precision table is based on the ranked list of documents using the similarity between each document and the query as produced by the same Sager & Lockemann's (1976) similarity measure.

Recall	Precision
at 0.25	1.00
at 0.50	1.00
at 0.75	0.98
at 1.00	0.38

Fig. 13: Recall-Precision based on 250 documents and the test query

By comparing precision values between Figures 12 and 13, it can be seen that precision at the 1.00 recall point has improved 4% (0.38 precision to 0.42 precision) from the traditional document-based information retrieval situation to the cluster-based retrieval. This is a surprising result in that clustering of documents has always been assumed to reduce information and negatively affect precision. An intellectual analysis of the clustered retrieval results shows that clustering effectively pulls up those individual documents which were ranked lowest in the document-based matching. However, the result reported here is suggestive, not definitive, as it is based on a small number of documents and a single query.

In terms of system efficiency, which is one of the major goals in cluster-based information retrieval, we reduced the retrieval computation by four-fifths by classifying 250 documents into 48 clusters. Since clustering will be done independently of queries, the document clusters can be used with all incoming queries.

CONCLUSION

Results of this preliminary experimentation cause us to be confident that the SFC vector representations are amenable to producing coherent, subject-based classifications using Ward's clustering algorithm. The SFC representation appears to offer sufficient specificity as well as allowing documents from different domains to be distinguished. Ward's minimum variance approach appears to produce clusters of internal coherence and does not suffer from some of the negative aspects of other clustering algorithms.

ACKNOWLEDGMENTS

We wish to thank Longman Group, Ltd. for making the machine readable version of LDOCE, 2nd Edition available to us and BBN for making POST available for our use on this project.

ISSN: 2324-9773

REFERENCES

Beckwith, R. & Miller, G.A. (1990). WORDNET Online lexical reference system.

Boguraev, B. & Briscoe, T. (1989). *Computational lexicography for natural language processin* London: Longman.

Choueka, Y. & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, pp. 147-157.

Dow Jones, (1989). *Wall Street Journal CD-ROM*.

Kelly, E. F. & Stone, P. J. (1975). *Computer recognition of English word senses*. Amsterdam: Nor Holland Publishing Co.

Krovetz, R. (1991). Lexical acquisition and information retrieval. In Zernik, U. (Ed.). *Lexical acquisition: exploiting on-line resources to build a lexicon*. Hillsdale, NJ: Lawrence Earlbaum.

Liddy, E.D. & Paik, W. (1992). Statistically-guided word sense disambiguation. In *Proceedings o AAAI Fall Symposium Series: Probabilistic approaches to natural language*. Menlo Par CA: AAAI.

Liddy, E.D. & Myaeng, S. H. (1991). Document Retrieval using LINguistic Knowledge (DR-LINK). Proposal in response to DARPA BAA#90-16.

Liddy, E.D. & Paik, W. (1991). An intelligent semantic relation assigner. *Proceedings of Worksho on Natural Language Learning*. Sponsored by IJCAI '91, Sydney, Australia.

Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the Fifteenth Annual International ACMSIGIR Conference o Research and Development in Information Retrieval*. Baltimore: ACM Press.

Meteer, M., Schwartz, R. & Weischedel, R. (1991). POST: Using probabilities in language processing. *Proceedings of the Twelfth International Conference on Artificial Intelligenc* Sydney, Australia.

Michalski, R.S. & Stepp, R.E. (1990). "Clustering". In S.C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (Vol. 1). New York: John Wiley & Sons.

Rasmussen, E. (1992). Clustering algorithms. In Frakes, W. & Baeza-Yates, R. (Eds.). *Informatio retrieval: Data structures and algorithms*. Englewood Cliffs, NJ: Prentice Hall.

SAS Institute Inc. (1988). *SAT/STAT (tm) User's guide*, Release 6.03 Edition. Cary, NC: SAS Institute Inc.

Sager, W.K.H. & Lockemann, P.C. (1976). Classification of ranking algorithms. *International Forum on Information and Documentation*. 1(4):2-25, 1976.

Slator, B. (1991). Using context for sense preference. In Zernik, U. (Ed.). *Lexical acquisition: exploiting on-line resources to build a lexicon*. Hillsdale, NJ: Lawrence Earlbaum.

Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.

Walker, D. E. & Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In Grishman, R. & Kittredge, R. (Eds). *Analyzing language in restricted domains Sublanguage description and processing*. Hillsdale, NJ: Lawrence Earlbaum.

Ward, J. (1963). Hierarchical grouping to optimize an objection function. *Journal of the Americar Statistical Association*. 58, p. 237-254.

Zupan, J. (1982). *Clustering of Large Data Sets*. New York: Research Studies Press.

ISSN: 2324-9773