STOCKHOLMS UNIVERSITET

Phrases or Words?

Clustering and Categorizing Swedish Scientific Medical Text

Sumithra Velupillai

TRITA-NA-E05063

Numerisk analys och datalogi
KTH
100 44 Stockholm

Department of Numerical Analysis
and Computer Science
Royal Institute of Technology
SE-100 44 Stockholm, Sweden

Phrases or Words?

Clustering and Categorizing Swedish Scientific Medical Text

Sumithra Velupillai

TRITA-NA-E05063

Master's Thesis in Computer Science (20 credits)
Single Subject Courses,
Stockholm University 2005
Supervisor at Nada was Viggo Kann
Examiner was Stefan Arnborg

Abstract

Clustering and categorization of documents is a large field within Information Retrieval and Natural Language Processing. Dealing with domain specific texts such as scientific medical text is fairly unexplored. This thesis investigates the possibilities of improving clustering results for Swedish scientific medical text by representing them with phrases instead of words. The idea is that phrases in this type of text carry more information than other types of texts. Newspaper articles are used as comparison material. Several phrase representations are presented. For the evaluation of the clustering results a reference partition where the articles are assigned predefined categories is required. Categorizations for the medical articles have been created using the MeSH-thesaurus. The hypothesis presented is not verified by the results. However, the results show that different representations of differing text types differ greatly.

Sammanfattning

Fraser eller ord?
Klustring och kategorisering av svensk medicinsk vetenskaplig text

Klustring och kategorisering av texter är ett stort område inom informationsextraktion och språkteknologi. Att hantera domänspecifika texter såsom medicinsk vetenskaplig text är någorlunda outforskat. I detta arbete undersöks det om man skulle kunna förbättra klustringsresultat av svensk medicinsk vetenskaplig text genom att istället för att representera texterna med extraherade ord, representera dem med substantivfraser. Tanken är att fraser i just denna typ av text är mer betydelsebärande än i annan text. Som jämförelsematerial används en korpus med nyhetsartiklar. Flera frasrepresentationer presenteras. För att evaluera klustringsresultaten krävs ett referensmaterial där artiklarna är indelade i fördefinierade kategorier. Kategoriseringar för de medicinska artiklarna har skapats med utgångspunkt i MeSH-tesauren. Hypotesen styrks inte av resultaten, däremot visas att olika representationer av olika texttyper skiljer sig markant åt.

Acknowledgments

I would like to thank some people for valuable help during the creation of this Master's thesis:

Magnus Rosell at Nada for providing and helping me with the clustering tool and for supervising and supporting me with ideas and comments on the work.

Prof. Viggo Kann for supervision and support.

Catharina Rehn, Jan-Erik Litton and everyone else at KI for providing me with MeSH-terms and ideas for the project.

Josef Milerad and everyone else at Läkartidningen (the Swedish Journal for Physicians) for providing me with the medical corpus and showing interest in the project.

Everyone involved with the Infomat-project.

Contents

Chapter 1

Introduction

This chapter gives an introduction to the subject presented in this thesis, including a background description, hypothesis and motivation, a small survey of the topics included and excluded in the project, as well as an outline of the report.

1.1 Background

Document clustering is an area within Information Retrieval, Text Mining and Natural Language Processing that has been studied for different purposes, such as improving information retrieval systems ([17]), browsing collections of documents ([3]), grouping retrieval results (see for example the search engine Vivisimo[1]). Clustering techniques are statistical methods used to find natural groupings in data.

However, most studies are based on experimenting and working with document collections in English ([8]). Studying and evaluating natural language processing applications for information retrieval purposes in other languages is both important and essential in order to enhance the systems and applications. Language specific tools have proved to be valuable when analyzing and working with large document collections. Studies on clustering Swedish texts have been performed, see for example [10] and [11], but many areas are left unexplored. The thesis presented here is part of a project called Infomat[2]. The purpose of Infomat is to explore topics in information retrieval for Swedish.

Document categorization is also an area within Information Retrieval and Natural Language Processing that has been studied for different purposes. Categorizing and classifying documents is essential when dealing with large document collections, organizing and sorting the collections makes the overview of the corpora easy and manageable. Many document collections have been manually categorized and indexed, a process that is very painstaking. If it is possible to create automatic systems that partly or completely substitute manual systems there is much to gain.

[1] http://vivisimo.com
[2] http://www.nada.kth.se/theory/projects/infomat/

Studying domain specific texts such as medical texts is both challenging and useful. Bioinformatics is a growing area where Natural Language Processing tools are incorporated and experimented with. The need for dealing with special vocabularies is large and important, for example in medical texts one might have to be able to handle complicated medical terms, gene expressions or other domain specific terms. Medical journals, article databases and libraries would profit a great deal if there were systems that handle medical texts in an efficient manner.

1.2 Hypothesis and Motivation

The hypothesis for this project is that clustering Swedish scientific medical text by representing it with noun phrases instead of words yields better results than if the same is done with other types of Swedish texts. The idea is that noun phrases in these domain specific texts carry more information than in other types of texts, since many medical terms are noun phrases and these should reflect the most important concepts. For other types of texts such as newspaper articles the thought is that although noun phrases may also carry important content information, these are not as characteristic for the content as for medical texts.

Much natural language processing work has been done with medical and biological text mining, but mostly for English. Since the vocabulary in medical texts differ from other types of texts different applications and systems would improve if it was possible to exploit domain specific characteristics. In particular it is necessary to work with these domain specific properties in Swedish, since it is yet left relatively uncharted. Although most scientific medical texts and articles are published in English in Sweden as well there are still journals and other media that publish work in Swedish, and some professions mainly utilize work written in Swedish. For that reason it would be interesting to see if clustering Swedish medical texts using phrase representations improves clustering results.

1.3 Topics Included and Excluded

There are many topics within information retrieval as well as document clustering and categorization that still awaits in depth analysis, many of which are related to the subject of this thesis. However, this project is restricted to cover only the questions included in the hypothesis. This means that several related topics are excluded in this project.

The thesis does not produce a survey on different document clustering algorithms or different ways of categorizing texts. To be more precise, the project studies the results of one special document clustering algorithm and creates categorizations for the medical texts that is not further explored. Discussions and evaluations of different ways of categorizing or clustering documents fall outside of the scope of this project. The measures used for the clustering process and for the evaluation of

the results are not further evaluated or discussed; these questions would be subject for further research.

The project does not evaluate the use of phrases instead of words for representing documents in general in information retrieval applications – rather, it studies the effects of making this distinction in the specific domain of medical text.

The project is about text analysis in Swedish, the results are not meant to be applied to other languages. However, the results might indicate that it could be interesting to look into such topics for other languages as well.

1.4 The Outline of this Master's Thesis

The thesis presented here is divided into four main parts. The first part, the current one, introduces the subject and hypothesis and gives a description of the purpose of the project. The second part (chapters 2–4) gives a survey on and description of the topics text categorization, document representations and linguistic features and finally document clustering. The third part (chapter 5 describes the steps taken in the project and the chosen methods and solutions). The final part (chapters 6–7) discusses the results and gives a discussion on future work and final thoughts.

Chapter 2

Text Categorization

In the following chapter a short introduction to the area of text categorization is given, as well as a description of MeSH and the function of MeSH in this project.

2.1 A Short Introduction to Text Categorization

When dealing with large document collections it is essential to organize them so that it is easy to extract information from them. Libraries, journals, archives, etc. need systems to arrange their data. Categories and classifications are developed and used for this purpose. There are many ways of deciding how to categorize items, much depending on the quality and attributes of what is to be categorized. Categorizations can be created with different purposes, and the assignment of an object to a specific category may differ from person to person or system to system. Nevertheless, categorizations are both important and useful for handling large corpora.

Text categorization is used in many differing fields, Yahoo for example has a web directory with numerous categories where sites are categorized[1], journals have different sections (categories) for articles, libraries organize their records in different category systems, search results from the internet are organized for example, to mention just a few areas of interest.

Defining the features one wishes to extract from texts in order to place them in one or several categories can be done in several ways. One way is to extract index terms from the text, choosing the terms one thinks describes the contents of the text in the most general yet accurate way. These index terms might come from a controlled vocabulary from a special domain (MeSH is an example of such a taxonomy (see section 2.2)) or they might be extracted in other ways, for example through an automated system.

Developing and maintaining taxonomies, category hierarchies, or other organization systems manually is a very costly, time consuming and exhaustive process. There are many studies on how to create systems of automatic text categorization ranging from automatic index term extraction to machine learning systems. For

[1] http://dir.yahoo.com/

more information on automatic text categorization applications and systems see the chapters on text categorization in [4] (chapter 4) or [8] (chapter 16). Chapter 14 in [16] describes the use of categories and clusters for organizing retrieval results.

For this study no automatic categorization systems were tested or evaluated. However, an already existing (mainly manually created) taxonomy (section 2.2) was used for the purpose of creating unambiguous categorizations of the articles in the medical corpus. These categorizations are later used when evaluating the clustering results of the corpora (see section 6 for discussions on evaluating the results).

Although clustering techniques should not be seen as categorization processes, there are similarities between the two: the function is to organize and divide the data into smaller, more structured parts that share similar characteristics. The significant difference between the methods is that clustering techniques find *structures already present in the data*, while categorization systems assign items to *predefined labels*.

2.2 MeSH: Medical Subject Headings

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus developed by the National Library of Medicine (NLM) in the United States for the purpose of indexing and cataloguing biomedical articles, journals and books. NLM are responsible for the production of MeSH. In Sweden the translations and indexation of Swedish biomedical articles, journals etc. are maintained by the library of Karolinska Institutet (KI). It is used in the MEDLINE/PubMED[2] database as well as the Swedish database SveMed+[3], produced by Karolinska Institutet, to name some examples. The MeSH thesaurus could be regarded as a tool for categorizing texts in a very detailed and meticulous way, in the sense that indexers assign index terms to all articles.

The terms in the thesaurus are organized both alphabetically and hierarchically. At the top level of the hierarchy the more general terms such as 'Anatomy', 'Enzymes' are, at the bottom level you find more specified terms such as 'Ankle', 'Myocardium'. The thesaurus is polyhierarchical, which means that a term can be found in several places of the hierarchy. However, each place in the hierarchy is unique and has a code assigned to it, which means that a term can be found in several places, but the places are corresponded by a distinct code. See figure 2.1 for an example from the MeSH hierarchy, where the polyhierarchic structure is exemplified. The term Multiple Sclerosis can be found in three places in the hierarchy, and each place has a unique code (i.e C.10.114.375.500, C10.314.350.500, C20.111.258.250.500).

There are 15 'top branches' of the hierarchy (see table 2.1). Each of these branches has a number of sub trees connected to them. In total there are about 22 500 descriptors/terms in the thesaurus. About 21 000 of these have Swedish translations.

[2]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3]http://micr.kib.ki.se/

MeSH Tree Structures

Nervous System Diseases [C10]
- Autoimmune Diseases of the Nervous System [C10.114]
  - Demyelinating Autoimmune Diseases, CNS [C10.114.375]
    - Diffuse Cerebral Sclerosis of Schilder [C10.114.375.112]
    - Encephalomyelitis, Acute Disseminated [C10.114.375.225]
    - Leukoencephalitis, Acute Hemorrhagic [C10.114.375.362]
    - Multiple Sclerosis [C10.114.375.500]
      - Multiple Sclerosis, Chronic Progressive [C10.114.375.500.200]
      - Multiple Sclerosis, Relapsing-Remitting [C10.114.375.500.600]
      - Neuromyelitis Optica [C10.114.375.500.650]
    - Myelitis, Transverse [C10.114.375.600] +
    - Neuromyelitis Optica [C10.114.375.650]

Nervous System Diseases [C10]
- Demyelinating Diseases [C10.314]
  - Demyelinating Autoimmune Diseases, CNS [C10.314.350]
    - Diffuse Cerebral Sclerosis of Schilder [C10.314.350.112]
    - Encephalomyelitis, Acute Disseminated [C10.314.350.225]
    - Encephalomyelitis, Experimental Autoimmune [C10.314.350.250]
    - Leukoencephalitis, Acute Hemorrhagic [C10.314.350.375]
    - Multiple Sclerosis [C10.314.350.500]
      - Multiple Sclerosis, Chronic Progressive [C10.314.350.500.200]
      - Multiple Sclerosis, Relapsing-Remitting [C10.314.350.500.600]
      - Neuromyelitis Optica [C10.314.350.500.650]
    - Myelitis, Transverse [C10.314.350.600] +
    - Neuromyelitis Optica [C10.314.350.650]

Immunologic Diseases [C20]
- Autoimmune Diseases [C20.111]
  - Autoimmune Diseases of the Nervous System [C20.111.258]
    - Demyelinating Autoimmune Diseases, CNS [C20.111.258.250]
      - Diffuse Cerebral Sclerosis of Schilder [C20.111.258.250.175]
      - Encephalomyelitis, Acute Disseminated [C20.111.258.250.350]
      - Leukoencephalitis, Acute Hemorrhagic [C20.111.258.250.425]
      - Multiple Sclerosis [C20.111.258.250.500]
        - Multiple Sclerosis, Chronic Progressive [C20.111.258.250.500.200]
        - Multiple Sclerosis, Relapsing-Remitting [C20.111.258.250.500.600]
        - Neuromyelitis Optica [C20.111.258.250.500.650]
      - Myelitis, Transverse [C20.111.258.250.550] +
      - Neuromyelitis Optica [C20.111.258.250.600]

Figure 2.1. An example from the MeSH thesaurus hierarchy.

Table 2.1. The MeSH Hierarchy: Top Branches.

A: Anatomy
B: Organisms
C: Diseases
D: Chemicals and Drugs
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment
F: Psychiatry and Psychology
G: Biological Sciences
H: Physical Sciences
I: Anthropology, Education, Sociology and Social Phenomena
J: Technology and Food and Beverages
K: Humanities
L: Information Science
M: Persons
N: Health Care
Z: Geographic Locations

All items indexed with MeSH terms and descriptors are indexed manually by a professional indexer. If clustering medical articles yields good results these could be used as a tool for indexers in order to speed up the indexation process, as an example of what use one could have from the results of this project.

More information about MeSH can be found at the official webpage for MeSH[4]. Information about the use of MeSH in Sweden can be found at the webpage for the library of Karolinska Institutet[5].

2.2.1 The Function of MeSH in this Project

For this project the MeSH thesaurus is used for categorizing the articles in the medical corpus. Each article in the corpus has been indexed with MeSH terms by indexers at Karolinska Institutet, both the original English terms and, where applicable, the Swedish translation. The MeSH terms assigned to each article can be regarded as a categorization on its own. However, these terms create an ambiguous and multi-dimensional categorization that can't be used when evaluating the clustering results, because each article ca only be assigned to one category in order for it to be possible to evaluate the performance of the clustering algorithm. For this reason the MeSH terms and their position in the hierarchy have been used to create four categorizations of the corpus (see section 5.2), where each article is assigned one and only one category. These categorizations are created in order to try to

[4]http://www.nlm.nih.gov/mesh/meshhome.html
[5]http://mesh.kib.ki.se/swemesh/about_se.cfm

flatten and disambiguate the category assignment for each article for the purpose of evaluating the clustering results of the corpora (see section 6).

Chapter 3

Document Representations and Linguistic Features

As stated above there are different ways of deciding in which way the contents of a document should be represented and extracted. For categorizing texts using the MeSH thesaurus for example, human indexers manually assign index terms and MeSH codes from the MeSH thesaurus so that they reflect the contents of the document in the best way possible. These index terms are used in the categorization process of the medical corpus (see section 5.2).

A common method for representing documents in both text clustering applications and information retrieval systems is the so called vector space model. This model is described in section 3.1. In order to reduce the representation space different linguistic features are employed, the ones used in this project are described in section 3.1.1. This chapter gives only a brief description of these issues, for deeper discussions see for example [2], [8], [4], [13].

3.1 The Vector Space Model

The vector space model is an approach for representing documents [13] and has been widely used in many information retrieval experiments.

Each document is represented by an n-dimensional vector, where n is the number of index terms in the document collection. These index terms could be words, phrases or other content identifiers.

The whole document collection is represented as a term-document matrix where each row of the matrix represents a document of n distinct terms t_1, t_2, ..., t_n, and each column represents the assignment of a specific term to the m documents d_1, d_2, ..., d_m of the collection. Every document-term combination is assigned a weight w_{ij} (see figure 3.1).

The assignment of a term is usually a weight, a real number, but it could also be for example a binary value. If weights are used the weight is supposed to reflect the importance of the corresponding term in the document.

$$\begin{bmatrix} w_{11} & w_{21} & \cdots & w_{i1} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{i2} & \cdots & w_{n2} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{1j} & w_{2j} & \cdots & w_{ij} & \cdots & w_{nj} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{1m} & w_{2m} & \cdots & w_{im} & \cdots & w_{nm} \end{bmatrix}$$

Figure 3.1. An example of a weighted term by document matrix.

In this model the terms are represented in a way that does not reflect the semantic connection between or order of the terms, rather, they are seen as a set where the terms aren't dependent on each other, the terms in a document are seen as a 'bag-of-words': a set of terms where the order of the terms is discounted, but the number of occurrences is recorded.

With this representation model it is possible to perform similarity calculations on the vectors using geometric interpretations of the distances of the different term vectors.

3.1.1 Linguistic Features

The purpose of the vector space model is to represent the contents of a document in the best possible way. All words in a document are not equally important when trying to capture the main content(s). When applying the vector space model on a corpus it is necessary to decide which terms should be regarded as more important than other terms.

Many words in a text don't contribute in capturing the main concepts of the document. In general linguistics it is common to distinguish between content words and function words. Content words are those which denote something in the 'real world' while function words serve content words in some way [9]. In the sentence 'The house has black windows' for example, 'house' and 'windows' denote something in the 'real world' while the word 'the' is a function word, it specifies one of the content words.

In order to reduce the sizes of the document vectors and better capture the contents of the documents it is common to sort out so called 'stop words'. These are usually function words as well as very frequent words (these do not contribute in making different documents distinguishable from each other). It is common to create stop word lists from large corpora and choosing the most frequent words and words that do not carry any information value. These are removed before creating the index matrices, and helps in reducing the sizes of the vectors.

In this project a Swedish stop word list was used to exclude stop words from both the newspaper corpus and the medical corpus, created from a large newspaper article corpus [10].

Lemmatization and Stemming

Many words are closely related and many words can be used in different forms. By converting these into one single form it is possible to reduce the document vector space considerably. There are two main ways of reducing different word forms to one (i.e. normalizing the terms): stemming and lemmatization. Stemming processes take away morphological and inflectional endings. The result is a 'basic' form: stem, this is not always corresponded by an actual word. The Swedish words 'fångst' (catch (noun)) and 'fånga' (capture (verb)) would yield the stem 'fång', which isn't an actual word. However, sometimes stemming might produce ambiguous stems: the words 'målning' (painting) and 'måla' (paint) would yield the stem 'mål' (goal, lawsuit, meal, dialect).

Lemmatization processes try to find the lemma or lexeme of inflected words: the lexeme is not an artificial word stem but an actual word: the basic, uninflected form of a word. For the examples given above the words 'fångst' and 'fånga' would yield the lemmas 'fångst' and 'fånga'. However, the words 'målning' and 'måla' would not yield ambiguous lemmas but rather the lemmas 'målning' and 'måla'.

The two processes are closely related, but stemming is more common in information retrieval contexts, much because there exist many stemming algorithms that perform reasonably well (they are both computationally effortless and they reduce the representation space considerably). However, as stated in the introduction, most of the information retrieval work so far has been about English. A study for Swedish [11] showed that both stemming and compound analysis improved clustering results. In this study both lemmatization and stemming on the corpora were employed, as well as compound analysis (see section 5.3).

Phrases and Compounds

A phrase is a semantically cohesive sequence of words [9]. A phrase is a syntactic structure which has syntactic properties derived from its head. In a noun phrase like 'the red car', 'car' is the head. For this study the focus lies in noun phrases, since most medical phrases are noun phrases, such as 'multiple sclerosis', 'cardiovascular disease'. As stated in the introduction the hypothesis is that these medical phrases differ in importance and carry more information about the content in the article, compared to phrases in 'ordinary' texts such as newspaper articles. In this project noun phrases were extracted using the tool 'Granska' (see section 5.3).

A compound is a lexeme that is composed of two or more roots [9]. Swedish is an inflectional language that builds words through combinations of different word roots. The noun 'alkoholkonsumtionsmönstret' (the pattern of alcohol consumption) contains the roots 'alkohol' (alcohol), 'konsumtion' (consumption) and 'mönster' (pattern) for example. Many medical terms are such compounds and might be seen as a phrase on their own. In order to find documents with similar contents it might be interesting to find documents that contain related terms such as 'alkoholkonsumtion' (alcohol consumption), and by splitting compounds to their different roots it might

be possible to find documents that share similar themes. Compound analysis was performed on the corpora (see section 5.3).

3.1.2 Weights and Similarity Measures

As stated above, in the term-document matrix each document-term combination is represented by a weight. There are different ways and theories of calculating these weights. In the area of information retrieval there are two general approaches to these weightings. One is concerned with the frequency of the term in the document (term frequency, tf):

$$tf_{ij} = \text{term frequency of term } i \text{ in document } j. \tag{3.1}$$

The other is concerned with the frequency of the term in the corpus as a whole: (inverse document frequency, idf):

$$idf_i = \text{inverse document frequency for term } i \text{ in document } j. \tag{3.2}$$

The final weight for a term i in document j in the document collection D is the product of the two weights:

$$weight_{ij} = tf_{ij} \cdot idf_i \tag{3.3}$$

Tf-idf-weighting is employed in the clustering tool used for this project. There are different ways of calculating these weights, alternative measures and alternative approaches to term weighting are not analyzed or evaluated in this project, see for example [10], [8], [4] for further discussions on calculating term weights.

Similarity Measure

The most common way to compute the similarity between documents, i.e. document vectors, in information retrieval contexts, is the cosine measure: the cosine of the angle between the document representation vectors. The smaller the angle, the more similar are the documents $d_1, d_2 \in D$. Note that the document vectors are normalized:

$$cosine(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \cdot \|d_2\|} = d_1 \bullet d_2 \tag{3.4}$$

$\bullet$ indicates the vector dot product, $\| d \|$ is the length of vector d.

3.1.3 Cluster Representation

Document clusters are also represented by term vectors. These cluster vectors depend on the term vectors of the document collection. This dependence can be

exploited in different ways, one common approach is to create centroids. The centroid vector c is the vector obtained by averaging the weights of the terms present in the document collection D:

$$c = \frac{1}{|D|} \sum_{d \in D} d \tag{3.5}$$

This approach is employed for this study, other alternatives are not examined but are left for further exploration in future research.

Chapter 4

Document Clustering

This chapter does not give a thorough description of clustering techniques in general nor does it give an in depth description of special features for the specific domain of document clustering. This chapter gives a short introduction to document clustering in general and a short discussion on the purpose of clustering documents. It also gives a description of the main clustering techniques commonly used for document clustering as well as the clustering algorithm used for this particular project. For more details concerning clustering in general (for areas outside natural language processing as well), see for example [6], [1].

Clustering is a statistical method which finds 'natural' groups in data [6]. Clustering techniques and applications are utilized in many different areas such as social sciences, biological sciences and text mining. As stated in the introduction, document clusters are used for organizing and browsing documents, document classification and for improving precision and recall in information retrieval systems, to mention a few areas of interest. If it is possible to yield good clustering results of a corpus consisting of medical texts these might be used as an aid for indexers who currently index manually using the MeSH thesaurus. In the future the automatic part might even prove to create such good results that the manual indexing part is but the final step in the classification process. This would save both time and money for many interested parties.

The general clustering process can be described in the following way (see chapter 6 (Document and Term Clustering) in: [7]):

1. Define domain.

2. Determine attributes of objects to be clustered.

3. Determine strength of relationships between attributes whose co-occurrence in objects suggests those objects should be in the same class, i.e. similarity function.

4. Apply algorithm to determine the class(es) to which each item will be assigned.

4.1 Clustering Techniques: Hierarchical and Partitional

Document clustering techniques are described in [8], [7] and [14], to name a few. The information in this section is primarily taken from these sources.

Hierarchical clustering techniques create a sequence of partitions that are nested, with singleton clusters of individual points at the bottom, and a single, all-inclusive cluster at the top. There are two basic methods to create hierarchical clusterings:

- agglomerative: all data points are considered individual clusters at the start, and at each step the two closest or most similar pair of clusters are merged
- divisive: one, all-inclusive cluster is used at the start, at each step a cluster is split until only singleton clusters of individual points remain.

Of the hierarchical clustering techniques, agglomerative methods are more common in document clustering applications. The results of hierarchical clusterings can be graphically displayed as a dendogram, i.e. a tree.

Partitional clustering techniques create a desired number of un-nested, one-level partitionings (clusters) of the data points. The most common method is the K-means algorithm, which is based on the idea that a center point can represent a cluster. The center point is usually a centroid (see section 3.1.3) (i.e. the mean or median of a group of points. In the clustering tool used for this project the mean is used).

It is possible to provide a hierarchical clustering by using a partitional clustering algorithm, and the same applies to hierarchical algorithms: it is possible to create flat partitions of the desired number of clusters. Bisecting K-means is a partitional algorithm that provides a hierarchical clustering (see section 4.2).

4.2 Bisecting K-means

The bisecting K-means algorithm was introduced in [14]. The details about the algorithm are quoted from the article. The algorithm starts with a single cluster containing all documents and works in the following way:

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. (Bisecting step)
3. Repeat step 2, the bisecting step for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

The cluster that is split is the largest remaining cluster, however there are other ways of choosing which cluster to split, for example one could choose the one with

the least overall similarity. In [14] they determined that the differences between the methods were small and they chose to split the largest remaining cluster.

The bisecting K-means technique proved to perform as good as or better than both the partitional K-means algorithm and the hierarchical techniques tested in the study by Steinbach et al. [14]. Bisecting K-means is used in the clustering tool used for this project, where the best split (using the similarity measure as evaluation) is picked.

4.3 Evaluating Clustering Results

Evaluating clustering results may be seen as a difficult task – how should one define a 'good' or adequate clustering result? How should one judge if one produced cluster is better or worse than another? What is a satisfactory partition for a set of objects? The criteria differ depending on what purpose a set is divided. These questions apply for the quality of categorizations too and characterizing and evaluating a clustering result or a categorization would probably yield different results if the task were given to different persons. These issues are subject for further debate and discussion.

Although evaluating clustering results is hard there are two main measures of deciding cluster qualities: internal quality measures and external quality measures. Internal quality measures compare different cluster sets without reference to external knowledge, for instance through similarity measures. External quality measures compare a clustering result with known classes from for example a reference categorization.

Internal quality measures have no function in this project since the focus lies in comparing different ways of representing documents using the same clustering algorithm, not evaluating different performances of different clustering algorithms.

In this project the categorizations created from the MeSH-terms (see section 5.2) are used as references for external quality measures. External quality measures are employed when evaluating the clustering results for the medical corpus and predefined categories for the newspaper corpus. The measures used for this project are described in the section below.

4.3.1 External Evaluation Measures

The definitions of and information about the measures described in this section are taken from [12], [14] and [15]. The measures used for evaluation are Entropy, Purity and Normalized Mutual Information.

In clustering contexts precision (p) is used: it compares each cluster i with each class j in the categorization:

$$p_{ij} = \frac{n_{ij}}{n_i} \tag{4.1}$$

where n_{ij} is the number of texts from class j in cluster i and n_i is the number of texts in cluster i.

Entropy

Entropy shows how the various classes of documents are distributed within each cluster. It could be called a measure of 'disorder'. In general a clustering solution is better the smaller the entropy value is.
The entropy for each cluster is

$$E_i = -\sum_j p_{ij} \log(p_{ij}) \tag{4.2}$$

and the weighted average entropy for the clustering as a whole is

$$E = \sum_i \frac{n_i}{n} E_i \tag{4.3}$$

where n is the number of texts in the whole document set.

Purity

Purity computes the extent to which each cluster contains documents from primarily one class. In general a clustering solution is better the larger the purity value is.
The Purity for each cluster is

$$\rho_i = \max_i \frac{n_{ij}}{n_i} \tag{4.4}$$

and the weighted average purity for the clustering as a whole is

$$\rho = \sum_i \frac{n_i}{n} \cdot \rho_i = \frac{n_{\max}}{n} \tag{4.5}$$

where $n_{\max}$ is the number of documents belonging to a cluster where its class is the largest.

Normalized Mutual Information

Mutual information is a symmetric measure to calculate the statistical information shared by two distributions. Normalized mutual information is the [0,1] normalized version of mutual information. The category Cat and the cluster C are treated as random variables, and the mutual information between them is:

$$MI(\text{Cat,C}) = \sum_i \sum_j \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_i n_j} \tag{4.6}$$

where n_j is the number of texts in category j.

In order to compute normalized mutual information one has to divide MI with max MI where

$$\max MI \leq \max \frac{E(Cat)+E(C)}{2} = \frac{\log(\kappa)+\log(\gamma)}{2} \tag{4.7}$$

where κ and γ are the number of categories/clusters.

Rand Statistic

The definitions of evaluation measures given above are all based on the distribution of single texts. It is also possible to utilize measures based on the distribution of pairs of texts. The Rand Statistic is the accuracy of pairwise decisions.

$$R = \frac{a+d}{m} \tag{4.8}$$

where $m = n(n-1)$ is the total number of pairs, a is the number of pairs in the same group in the first partition and in the second, and d is the number of pairs in different groups in the first partition and in the second.

Chapter 5

Project Outline: Methods and Solutions

In order to test the hypothesis stated in the introduction several steps had to be taken. A corpus had to be collected and prepared, the categorizations used for the evaluation process had to be created, the texts had to be manipulated in order to extract the desired words and phrases, and finally the corpora had to be clustered. An outline of the project is shown in Figure 5.1, and the steps taken are further described in the sections below.

5.1 The Corpus: Collection and Preparation

The Swedish journal for physicians, Läkartidningen[1], is a journal that publishes both scientific articles and popular science articles in the domain of medicine. Their article database contains only files in the pdf-format, for this reason all articles were converted to plain text using the tool pdf2txt. Articles shorter than one page and articles that did not contain any references were excluded in order to capture only scientific articles. This resulted in a corpus consisting of 2422 articles, with a total of 4 383 169 (26 102 different) words.

The corpus used as a contrasting text type is a corpus consisting of newspaper articles from the KTH News Corpus. It is a subset of the large corpus consisting of 2500 articles, with a total of 119 401 (5896 different) words. These articles are already categorized in categories such as Economy and Sports.

5.2 Categorizing the Medical Corpus

All articles are part of the database SveMed+ (produced and maintained by the library at Karolinska Institutet). Each article has been indexed with a number of MeSH-terms. These index terms (for each article) were extracted in order to be used in the creation of a categorization for the corpus.

As stated above, the process of categorizing text is not always unproblematic. There are always issues that could be debated when assigning a text to a predefined

[1] http://www.lakartidningen.se/

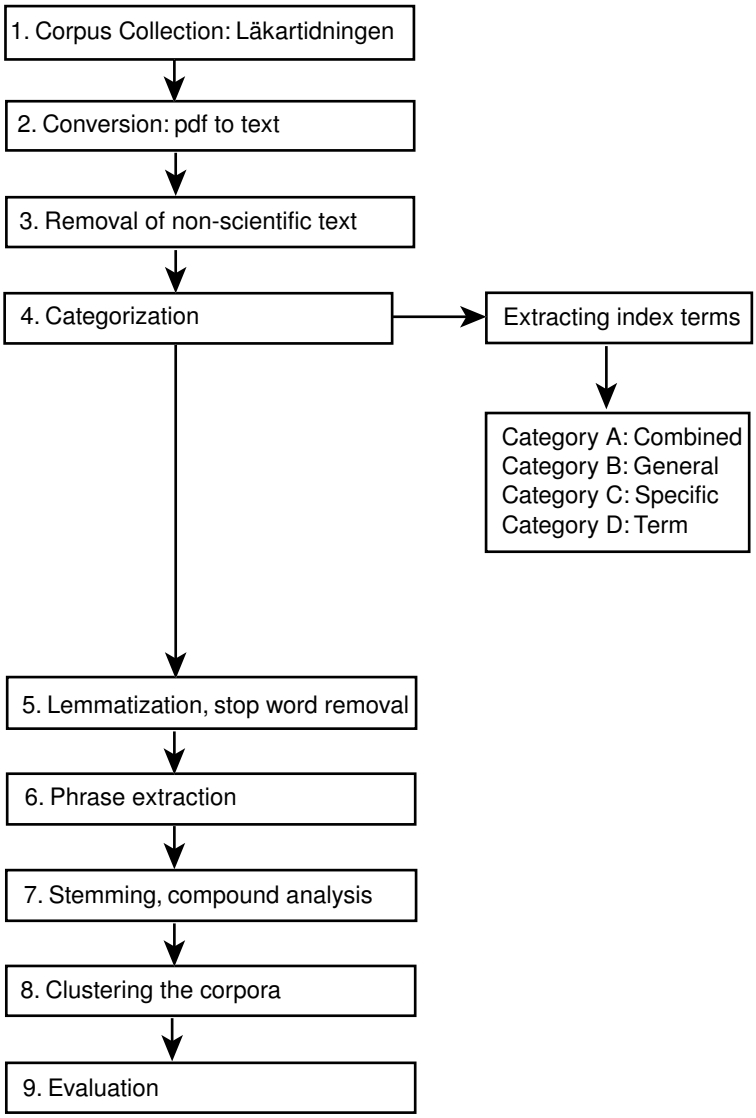

Figure 5.1. Project Outline.

Title:
Immunsvar och inflammation vid Crohns sjukdom.
Snart möjligt driva diagnostiken längre och utveckla mer specifika läkemedel

MeSH Heading: Crohn disease/DI/IM/DT/GE

MeSH codes:
C06.405.469.220 (Crohn Disease (Crohns sjukdom)
C06.405.469.440.148 (Crohn Disease (Crohns sjukdom))
A11.118.637.555.567.569.200.400 (T-Lymphocytes, Helper-Inducer (Hjälpar-T-celler))
A15.145.229.637.555.567.569.200.400 (T-Lymphocytes, Helper-Inducer (Hjälpar-T-celler))
A15.382.490.555.567.569.200.400 (T-Lymphocytes, Helper-Inducer (Hjälpar-T-celler))
A15.382.520.520.604.200.400 (T-Lymphocytes, Helper-Inducer (Hjälpar-T-celler))

Figure 5.2. An example file from Läkartidningen 1999, article no: 18827.

category, there will always be texts that are difficult to assign one single category. In addition, assigning an article to one and only one category is even more difficult when the contents might reflect equal membership in several categories. An article about 'Multiple Sclerosis' might also have as its main content 'Health Care'. The advantages of the MeSH-indexing process is that it is possible to assign many index terms to an article, which means that an article can be assigned to many different categories.

In this project it was necessary to create a categorization (step 4 in the process, see figure 5.1) that does not allow an article to be part of several categories, because when evaluating the results it is not possible to allow an item to be part of more than one cluster. For the purpose of evaluating the clustering results a categorization where each article is assigned one and only one category was created, using the MeSH terms (the unique codes) that each article was indexed with (see figure 5.2 for an example of the index terms for an article (only the title is given, not the whole text)). We tried to create categories that were general enough to contain documents with similar contents and specific enough to capture the main content of the single documents.

Since the MeSH index system is very extensive and each indexer tries to find as many relevant index terms as possible for each article, several of the index terms might not be as good in capturing the main content of the article as others. Choosing one of these terms for the automatic creation of a category system was not easy, which is why four category types were created, utilizing different features of the MeSH terms. The category types are described below.

5.2.1 Head Aspect (Combined)

For some articles (but not all) a special index term called Mesh Heading was used to indicate the main aspects of the article. If the article had one Mesh Heading, this was used for the category 'Head Aspect'. If the article had more than one head

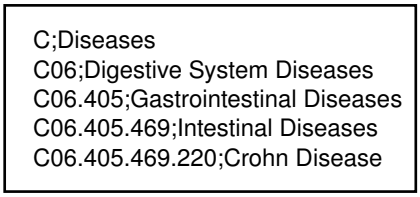

Figure 5.3. The Combined category assignment for the example file in figure 5.2.

aspect, only one was chosen, always the first. The motivation for this was that creating a system that automatically decided which one would be most appropriate would be too demanding. The term and all its parent nodes were used to place the article in a corresponding tree structure. If the article was not indexed with a Mesh Heading, the result from the Specific categorization (see section 5.2.3) was used – for this reason the category type is called 'Combined'. In figure 5.3 the result of the Combined categorization for the file in figure 5.2 is showed. The total number of files for the category (top level) is shown in Appendix A, table A.1.

5.2.2 General

Each MeSH term is a part of one of the 15 top branches of the MeSH hierarchy (figure 2.1). One of the created category types utilizes these to make a 'broad' categorization. The top branch of each index term is saved and counted, and the one that most index terms belong to is the category the article is assigned to. All articles are in this way assigned to one of the 15 top branches. Some articles have equally many index terms assigned to several top branches, in this case the first one is chosen. In the example (figure 5.2) the article is assigned to the category 'A: Anatomy'. The total number of files for the category is shown in Appendix A, table A.2.

5.2.3 Specific

One idea for deciding a category type was to use the most 'specific' index term for an article, i.e. the term that is situated 'deepest' in the hierarchy. The longest index term was chosen (if several terms were equally long, the one from the most frequent top branch was chosen). All parent branches were extracted and the file was assigned all sub-categories (see figure 5.4). The total number of files for the category (top level) is shown in Appendix A, table A.3.

5.2.4 Term

The MeSH codes are corresponded by a term in the MeSH thesaurus. As stated above, each article has been indexed with a number of terms (MeSH codes), since the thesaurus is polyhierarchical, for some files many index codes actually represent the same term. This is captured in the category 'Term': the corresponding term

A;Anatomy
A15;Hemic and Immune Systems
A15.145;Blood
A15.145.229;Blood Cells
A15.145.229.637;Leukocytes
A15.145.229.637.555;Leukocytes, Mononuclear
A15.145.229.637.555.567;Lymphocytes
A15.145.229.637.555.567.569;T-Lymphocytes
A15.145.229.637.555.567.569.200;CD4-Positive T-Lymphocytes
A15.145.229.637.555.567.569.200.400;T-Lymphocytes, Helper-Inducer

Figure 5.4. The Specific category assignment for the example file in figure 5.2.

for each MeSH code is found, and the most frequent term is chosen (in the example file shown in figure 5.2 the Term category is 'T-Lymphocytes, Helper-Inducer'). For some files no term is more frequent than another, in that case the first is chosen. This categorization yields many different categories, where some only contain one file, which makes it the most unstable categorization to use for the evaluation.

5.3 Document Representation: Lemmatization and Phrase Extraction

To split each article into lexemes and phrases a tool called Granska[2] was used. This tool was developed at Nada for the purpose of having a grammar checking tool for Swedish. The tool extracts lexemes as well as syntactic information such as part-of-speech tags and phrase information.

The lemmatization process was quite straightforward, this information was extracted easily. Granska produces several types of outputs when analyzing a text, one of these outputs is lemma information. However, when studying the results after lemmatizing we noticed that many words (mostly medical terms) hadn't been lemmatized properly. For example, the words 'amyloidens' (the amyloid's) and 'amyloiden' (the amyloid) were not lemmatized to one form, which is why we decided to process the lemmas again with a stemming algorithm [5], which would yield the stem 'amyloid'.

Swedish is an inflectional language, where long and intricate compounds can be found, for example 'amyloidprekursorproteingenen' (the amyloid precursor protein gene) (noun). These might function as (noun) phrases on their own, and be considered as such in the representation, which is why we have chosen to experiment with treating compounds as phrases in some of our representations.

The phrase extraction process required some decisions about which phrases to choose. The medical phrases are primarily noun phrases, and all noun phrases from

[2]http://www.nada.kth.se/theory/projects/granska/

the corpora were extracted using Granska. After extracting the phrases we discovered on one hand that many phrases were not longer than one word (i.e. single nouns or compounds), on the other hand that many of the phrases containing more than one word consisted of words that could be regarded as stop words. The following phrases were extracted for example: 'ökad kardiovaskulär sjukdom' (increased cardiovascular disease) and 'svår kardiovaskulär sjukdom' (severe cardiovascular disease). We decided to remove stop words from the phrases in order to extract better, 'purer' phrases. We also created a trie of the words in the phrases in order to find overlaps of words and phrases in the different documents.

The usual representation used in information extraction contexts is a stemmed word based vector space model. For this project this representation is used as a reference. A second representation is one where split compounds are inserted, but combined with words, not phrases.

The representations based on phrases are first divided into two types: one where the similarities are calculated only with phrases, one where the similarities are calculated both with the phrases and the words. Two ways of regarding phrases as similar are employed: either if the phrases are identical or if they share common words within the phrases, i.e. if there is an overlap between the words in the phrases. In the first case we have calculated the weight for each phrase in a document as the frequency of its appearance in that document multiplied by the sum of the idf-weight for the single words in it. In the second case we have built a trie based on words for each document from the phrases appearing in them. Each phrase is put into the trie in its entire and with all but the first word, with all but the first two words, etc. In each node of the trie we save the number of times it has been reached. To calculate the overlap of phrases between two documents we follow all common paths in the tries and multiply relative appearances in each node weighted by the sum of the idf-weights for the words along the path. We have also chosen to investigate the implications of regarding compounds as phrases or not as well as looking at the impacts of splitting compounds within phrases or not. All in all 18 representations were employed: eight phrase representations (full phrases or overlap, split compounds or not and split compounds within phrases or not), eight word representations (with the same possible combinations) as well as the reference representation (only words) and the representation where solid compounds have been split.

Table B.1 in Appendix B contains a summary of the representations employed.

5.4 Clustering the Corpora

For the clustering process a tool developed by Magnus Rosell at Nada was used. This clustering tool employs the clustering algorithm Bisecting K-means (see section 4.2). All documents in the medical corpus (2422 articles in total) as well as all articles in one of the newspaper corpora (2500 articles in total) were clustered, and all documents were clustered with each of the representations described above.

The K-Means algorithm was iterated ten times, for each split it was run five times picking the best split (using the similarity measure as evaluation). All average results were calculated over ten runs to ten clusters for each representation. The clustering results are discussed in the following chapter. All result tables are shown in Appendix B.

Chapter 6

Results and Evaluation

In Appendix B the result tables from the clusterings are shown. The results are unfortunately not encouraging. In general it is shown that using phrases for the representation does not increase the performance of the clustering processes, instead they generate much worse results in some cases. However there is a tendency that treating compounds as phrases does yield good results, at least for the newspaper articles, but perhaps they should be extracted, used or manipulated in some other way. The most discouraging part of the results is that the hypothesis that using phrases for the representations of the medical texts would yield better results is contradicted, the most interesting results are seen in the newspaper article corpus. One aspect of the outcome that is worthy of note is that the results show that different representations diverge a lot between the corpora, which indicates that it could be useful to treat different text types in specialized ways.

6.1 The Categorizations

The performances of the created categorizations are somewhat differing. Two of the categorizations created for this project seem to be the most trustworthy and coherent when looking at the results: Combined and General. The categorization called Term generated many categories, where several only contain one article, and many of them are found at different levels of the MeSH-hierarchy, which probably yields a disproportionate and uneven categorization. When using specific terms these do not necessarily reflect the head aspect, which might result in the assigning of an article to a category that does not reflect the general content of the article. The categorizations Combined and General probably work better because each category is more extensive and generally descriptive. However, as discussed earlier, it is difficult to assess whether the categorizations are 'good' or 'bad', for future research in this area it would be interesting to evaluate the categorizations too to see if they could have been 'better' in some aspect.

6.2 The Cluster Results

Table B.1 explains which representation is used for which clustering result. Tables B–B show the clustering results for the medical corpus, tables B–B show the clustering results for the newspaper article corpus, with the values for the evaluation measures Entropy, Purity, Normalized mutual information and Rand statistic (with the variance in parenthesis) and the difference in percentage with respect to the reference representation (number 1, tables B and B respectively, in Appendix B).

For both corpora representation 2 performs the best, i.e. the representation with stemmed words and split compounds. For the medical corpus the entropy decreases with 0.3 %, NMI increases with 2 % (for the category Combined), for the newspaper corpus the entropy decreases with 10.6 %, NMI increases with 15.4 %.

6.2.1 The Medical Corpus

The results for the medical corpus are as stated before not encouraging. All results are better than the random clustering, which is good, but the phrase representations do not perform better than the reference representation or the representation with stemmed words and split compounds. Representation 18 works the best (6.8 % entropy increase, 48.1 % NMI decrease for Combined for example), which still is a very poor result.

There is a tendency for the representations where only phrases are used (i.e. 3–10, tables B and B) that considering compounds as phrases for these representations does not improve any results, on the contrary, these representations yield worse results (for Combined, NMI for representations 5–6 and 9–10 lies around -92,5 % while representations 3–4 and 7–8 lies from -60,6 % to -74,5 %). A difference between the treatment of the two corpora is, as stated earlier, that a list of words that should not be split when performing compound analysis was created for the newspaper corpus, but no such domain specific list was created for the medical corpus. Creating such a word list might yield different results, where the differences might be the most clear for the representations mentioned.

It might be the case that the created categories aren't satisfactory. Perhaps these should have been extracted in a completely different way. The purpose of the MeSH-terms is that it is possible to assign both very detailed and general index terms and these together form a very precise 'categorization' of the articles. Maybe it is not possible to summarize them in the way we did when forming the categories.

The text type is, as discussed earlier, very specific, which might make it more difficult to use for clustering. Maybe it is more difficult to make adequate representations of such texts. In this case the texts are also much larger than the newspaper articles, which might influence the outcome.

An interesting topic when dealing with medical texts is the treatment and extraction of synonyms. Many medical terms have several synonyms; it would be interesting to see if they are used extensively in articles such as the ones in this

corpus. If it is common, there would be many terms that could be projected to one form, which probably would yield more representative representations.

Perhaps it is more difficult for this text type than for text types such as shorter newspaper articles to assign documents only one category. It would be interesting to look at the use of category and clustering hierarchies for this domain specific text type, maybe the results would differ a great deal.

6.2.2 The Newspaper Corpus

For the newspaper corpus the results were more interesting. The differences between the different representations were very large in some cases. The representations 3–10 (tables B, B) perform very badly, almost as bad as a random clustering.

However, the representations 11–18 (tables B, B) differ a lot, where the representations 13–14, 17–18 perform almost as good as the reference representation. These are the representations where phrases and words are used (PW) and compounds are regarded as phrases (CP), with the difference that real phrases are used in 13 and 14, and phrase-tries are used in 17–18. It would be very interesting to try to enhance the performance of these representations, perhaps by trying to improve the use of treating compounds as phrases.

6.2.3 General Thoughts

It might be the case that the results differ so greatly because the reference partitions, i.e. the categorizations, are very dissimilar. It is difficult to compare the two corpora because of this divergence, but still there are aspects which makes it interesting to compare the results. The newspaper articles are in general much shorter than the medical articles and the predefined categories are more defined, which might influence the performance of the clusterings and the representations. However, the results for the medical texts still do suggest that it might be better to focus on another approach for trying to improve clustering results at least in Swedish.

Chapter 7

Future Work and Final Remarks

Although the results were not as positive as one might have hoped for they do open doors for further thoughts and ideas about how to improve, simplify and study the processes of categorizations and document clustering techniques.

Dealing with medical texts seems to differ noticeably from dealing with other types of texts, in this case news paper articles. Maybe it would be interesting to see how results could improve if it was possible to work with synonyms. Many medical terms can be used in different ways although they reflect the same concepts (this is clearly shown in the MeSH-thesaurus) – if it was possible to project the same concepts to one and only one concepts the representations created would both be smaller and reflect the concepts in a more accurate and unambiguous way. Perhaps the use of a medical ontology would be useful for example.

Perhaps it could be interesting to work further with the creation of the representations of the documents. The representations constructed for this project might not be faultless. If continuing doing research in this area it would be interesting to work and experiment more with the representations of the documents. One aspect that should be taken into account is that when dealing with the compounds in this project a special list of words that should NOT be split was used, which proved to be useful [10]. However, this list was created from the newspaper corpus. A similar list for medical terms might have affected the results. Such a domain specific list would be interesting to create if this subject was investigated further.

This project has focused on the impacts of using phrases for the representations of Swedish texts. The results do not show how the same study would perform if analysing texts in other languages. English for example is a language where phrases are employed and created somewhat differently, maybe the results would be better for such languages, or other languages where phrases are not utilized in the same way as in Swedish.

Assigning an article just one category or cluster might be debatable in many contexts – most texts have more than one main subject matter. It might be interesting to see how clustering techniques where it is possible to assign a document more than one cluster would work.

Even though the outcome of the hypothesis proposed for this project was not affirmative many interesting subjects and questions for further research have emerged and only the future will reveal how far one could investigate these issues.

References

[1] P. Arabie, L. J. Hubert, and G. De Soete. *Clustering and Classification.* World Scientific Publishing Co. Pte. Ltd., 1996.

[2] M. Berry, Z. Drmac, and E. R Jessup. Matrices, vector spaces, and information retrieval. In *SIAM Rev. 41 (1999) no. 2, pages 335–362 (electronic)*, http://epubs.siam.org/sam-bin/dbq/article/34703 (last visited 050410), 1999.

[3] D. R. Cutting, J. O. Pedersen, D. Karger, and Tukey J. W. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1992.

[4] P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization.* Natural Language Processing. John Benjamins Publishing Company, 2002.

[5] V. Kann, R. Domeij, J. Hollman, and M. Tillenius. Implementation aspects and applications of a spelling correction algorithm. *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, 60, 2001.

[6] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data.* John Wiley & Sons, Inc., 1990.

[7] G. Kowalski. *Information Retrieval Systems – Theory and Implementation.* Kluwer Academic Publishers, 1997.

[8] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, 2001.

[9] S. C. Poole. *An Introduction to Linguistics.* St. Martins Press, 1999.

[10] M. Rosell. *Klustring av svenska tidningsartiklar.* Master's thesis, (In Swedish), Kungliga Tekniska Högskolan, 2002.

[11] M. Rosell. Improving clustering of swedish newspaper articles using stemming and compound splitting. In *Proc. 5th Nordic Conf. on Comp. Ling. – NODALIDA '03*, 2003.

[12] M. Rosell, V. Kann, and J. E Litton. Comparing comparisons: Document clustering evaluation using two manual classifications. In *ICON*, 2004.

[13] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1983.

[14] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proc. Workshop on Text Mining, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2000.

[15] A. Strehl and J. Ghosh. Cluster ensembles a knowledge reuse framework for combining partitionings. In *Proc. of AAAI, Edmonton, Canada*, 2002.

[16] T. Strzalkowski, editor. *Natural Language Information Retrieval.* Kluwer Academic Publishers, 1999.

[17] C. J van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.

Appendix A

Category Tables

Table A.1. Combined: number of files, top level.

Category	Nr of files
A: Anatomy	45
B: Organisms	32
C: Diseases	322
D: Chemicals and Drugs	183
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment	446
F: Psychiatry and Psychology	255
G: Biological Sciences	458
H: Physical Sciences	30
I: Anthropology, Education, Sociology and Social Phenomena	164
J: Technology and Food and Beverages	29
K: Humanities	94
L: Information Science	57
M: Persons	37
N: Health Care	260
Z: Geographic Locations	10

Table A.2. General: number of files, top level.

Category	Nr of files
A: Anatomy	22
B: Organisms	14
C: Diseases	472
D: Chemicals and Drugs	279
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment	393
F: Psychiatry and Psychology	122
G: Biological Sciences	265
H: Physical Sciences	10
I: Anthropology, Education, Sociology and Social Phenomena	62
J: Technology and Food and Beverages	3
K: Humanities	42
L: Information Science	41
M: Persons	44
N: Health Care	634
Z: Geographic Locations	19

Table A.3. Specific: number of files, top level.

Category	Nr of files
A: Anatomy	49
B: Organisms	44
C: Diseases	328
D: Chemicals and Drugs	248
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment	227
F: Psychiatry and Psychology	133
G: Biological Sciences	713
H: Physical Sciences	38
I: Anthropology, Education, Sociology and Social Phenomena	178
J: Technology and Food and Beverages	6
K: Humanities	31
L: Information Science	84
M: Persons	30
N: Health Care	279
Z: Geographic Locations	34

Appendix B

Clustering Tables

Table B.1. Abbreviations.

Abbreviation	Explanation
P	Phrase similarity only
P&W	Phrase and word similarity
RP	Real phrases
PTP	PhraseTrie-phrases
CP	Compounds regarded as phrases
CNP	Compounds NOT regarded as phrases
CSP	Compounds split within phrases
CNSP	Compounds NOT split within phrases

Table B.2. Representations.

Representation				
Tot	'Clustering', containing one cluster			
Rand	Random partition of the set, the average of ten parts			
1	stemming, no compound analysis			
2	stemming and compound analysis			
3	P	RP	CNP	CNSP
4	P	RP	CNP	CSP
5	P	RP	CP	CNSP
6	P	RP	CP	CSP
7	P	PTP	CNP	CNSP
8	P	PTP	CNP	CSP
9	P	PTP	CP	CNSP
10	P	PTP	CP	CSP
11	P&W	RP	CNP	CNSP
12	P&W	RP	CNP	CSP
13	P&W	RP	CP	CNSP
14	P&W	RP	CP	CSP
15	P&W	PTP	CNP	CNSP
16	P&W	PTP	CNP	CSP
17	P&W	PTP	CP	CNSP
18	P&W	PTP	CP	CSP

Table B.3. Text Clustering with Phrases: Medical Corpus: Tot, Rand, 1–2.

	Measures	Tot		Rand		1		2	
Combined	E	3.301	(0.000)	3.262	(0.004)	2.894	(0.016)	2.886	(0.010)
		14.1%	(0.0%)	12.7%	(0.1%)	0.0%	(0.6%)	−0.3%	(0.3%)
	NMI	0.000	(0.000)	0.011	(0.001)	0.113	(0.004)	0.115	(0.003)
		−100.0%	(0.0%)	−90.4%	(1.0%)	0.0%	(4.0%)	2.0%	(2.4%)
	ρ	0.189	(0.000)	0.201	(0.004)	0.281	(0.011)	0.289	(0.005)
		−32.7%	(0.0%)	−28.6%	(1.4%)	0.0%	(3.9%)	2.8%	(1.7%)
	R	0.066	(0.000)	0.012	(0.000)	0.808	(0.001)	0.807	(0.001)
		−91.8%	(0.0%)	−98.5%	(0.0%)	0.0%	(0.1%)	0.0%	(0.2%)
General	E	2.963	(0.000)	2.927	(0.003)	2.485	(0.013)	2.476	(0.016)
		19.2%	(0.0%)	17.8%	(0.1%)	0.0%	(0.5%)	−0.3%	(0.6%)
	NMI	0.000	(0.000)	0.010	(0.001)	0.132	(0.004)	0.135	(0.004)
		−100.0%	(0.0%)	−92.5%	(0.7%)	0.0%	(2.7%)	1.7%	(3.4%)
	ρ	0.262	(0.000)	0.262	(0.001)	0.388	(0.005)	0.391	(0.007)
		−32.5%	(0.0%)	−32.5%	(0.1%)	0.0%	(1.3%)	0.7%	(1.7%)
	R	0.066	(0.000)	0.016	(0.000)	0.785	(0.001)	0.785	(0.001)
		−91.6%	(0.0%)	−97.9%	(0.0%)	0.0%	(0.2%)	0.1%	(0.2%)
Specific	E	3.181	(0.000)	3.141	(0.005)	2.784	(0.010)	2.784	(0.017)
		14.2%	(0.0%)	12.8%	(0.2%)	0.0%	(0.4%)	0.0%	(0.6%)
	NMI	0.000	(0.000)	0.011	(0.001)	0.110	(0.003)	0.110	(0.005)
		−100.0%	(0.0%)	−90.1%	(1.2%)	0.0%	(2.6%)	0.1%	(4.2%)
	ρ	0.294	(0.000)	0.294	(0.000)	0.319	(0.009)	0.316	(0.011)
		−7.8%	(0.0%)	−7.8%	(0.0%)	0.0%	(3.0%)	−1.0%	(3.4%)
	R	0.066	(0.000)	0.015	(0.000)	0.786	(0.001)	0.786	(0.002)
		−91.6%	(0.0%)	−98.1%	(0.0%)	0.0%	(0.2%)	−0.1%	(0.2%)
Term	E	8.241	(0.000)	6.788	(0.008)	6.391	(0.023)	6.373	(0.018)
		28.9%	(0.0%)	6.2%	(0.1%)	0.0%	(0.4%)	−0.3%	(0.3%)
	NMI	0.000	(0.000)	0.224	(0.001)	0.285	(0.004)	0.287	(0.003)
		−100.0%	(0.0%)	−21.5%	(0.4%)	0.0%	(1.2%)	1.0%	(0.9%)
	ρ	0.105	(0.000)	0.105	(0.000)	0.121	(0.004)	0.119	(0.005)
		−13.0%	(0.0%)	−13.0%	(0.0%)	0.0%	(3.4%)	−0.9%	(4.5%)
	R	0.001	(0.000)	0.002	(0.000)	0.883	(0.002)	0.882	(0.002)
		−99.9%	(0.0%)	−99.8%	(0.0%)	0.0%	(0.2%)	−0.1%	(0.3%)

Table B.4. Text Clustering with Phrases: Medical Corpus: 3–6.

	Measures	3		4		5		6	
Combined	E	3.202	(0.007)	3.198	(0.007)	3.280	(0.007)	3.278	(0.003)
		10.6%	(0.3%)	10.5%	(0.2%)	13.3%	(0.2%)	13.3%	(0.1%)
	NMI	0.027	(0.002)	0.029	(0.002)	0.008	(0.002)	0.009	(0.001)
		−75.6%	(1.8%)	−74.5%	(1.7%)	−92.7%	(1.7%)	−92.3%	(0.8%)
	ρ	0.212	(0.004)	0.211	(0.006)	0.199	(0.001)	0.200	(0.001)
		−24.4%	(1.3%)	−25.0%	(2.1%)	−29.2%	(0.5%)	−28.8%	(0.4%)
	R	0.797	(0.001)	0.798	(0.002)	0.548	(0.038)	0.550	(0.020)
		−1.3%	(0.1%)	−1.2%	(0.3%)	−32.2%	(4.7%)	−31.9%	(2.5%)
General	E	2.851	(0.008)	2.845	(0.012)	2.942	(0.005)	2.939	(0.002)
		14.7%	(0.3%)	14.5%	(0.5%)	18.4%	(0.2%)	18.3%	(0.1%)
	NMI	0.031	(0.002)	0.032	(0.003)	0.008	(0.001)	0.008	(0.001)
		−76.6%	(1.7%)	−75.4%	(2.5%)	−94.2%	(1.1%)	−93.7%	(0.4%)
	ρ	0.293	(0.006)	0.294	(0.009)	0.264	(0.000)	0.265	(0.000)
		−24.4%	(1.6%)	−24.2%	(2.3%)	−31.8%	(0.1%)	−31.8%	(0.1%)
	R	0.768	(0.001)	0.769	(0.002)	0.543	(0.034)	0.544	(0.018)
		−2.1%	(0.1%)	−2.0%	(0.3%)	−30.9%	(4.3%)	−30.6%	(2.3%)
Specific	E	3.083	(0.008)	3.078	(0.008)	3.156	(0.007)	3.155	(0.004)
		10.7%	(0.3%)	10.6%	(0.3%)	13.4%	(0.3%)	13.3%	(0.1%)
	NMI	0.027	(0.002)	0.028	(0.002)	0.009	(0.002)	0.009	(0.001)
		−75.4%	(1.9%)	−74.2%	(2.0%)	−91.9%	(1.8%)	−91.8%	(1.0%)
	ρ	0.295	(0.001)	0.295	(0.001)	0.297	(0.000)	0.297	(0.001)
		−7.7%	(0.3%)	−7.8%	(0.2%)	−7.0%	(0.1%)	−6.9%	(0.2%)
	R	0.777	(0.001)	0.778	(0.002)	0.543	(0.035)	0.545	(0.018)
		−1.1%	(0.1%)	−1.1%	(0.3%)	−30.9%	(4.4%)	−30.6%	(2.3%)
Term	E	6.694	(0.020)	6.694	(0.013)	7.687	(0.097)	7.675	(0.023)
		4.7%	(0.3%)	4.7%	(0.2%)	20.3%	(1.5%)	20.1%	(0.4%)
	NMI	0.238	(0.003)	0.238	(0.002)	0.088	(0.015)	0.090	(0.003)
		−16.4%	(1.1%)	−16.4%	(0.7%)	−69.0%	(5.2%)	−68.4%	(1.2%)
	ρ	0.110	(0.003)	0.107	(0.001)	0.107	(0.000)	0.107	(0.000)
		−9.1%	(2.3%)	−11.2%	(1.0%)	−10.9%	(0.3%)	−11.0%	(0.3%)
	R	0.880	(0.001)	0.881	(0.003)	0.559	(0.049)	0.562	(0.027)
		−0.3%	(0.2%)	−0.3%	(0.3%)	−36.7%	(5.6%)	−36.4%	(3.0%)

Table B.5. Text Clustering with Phrases: Medical Corpus: 7–10.

	Measures	7		8		9		10	
Combined	E	3.166	(0.010)	3.141	(0.017)	3.281	(0.004)	3.278	(0.017)
		9.4%	(0.4%)	8.5%	(0.6%)	13.4%	(0.1%)	13.3%	(0.6%)
	NMI	0.037	(0.003)	0.044	(0.005)	0.008	(0.001)	0.009	(0.003)
		−66.8%	(2.6%)	−60.6%	(4.3%)	−92.7%	(0.6%)	−92.3%	(3.1%)
	ρ	0.212	(0.004)	0.213	(0.005)	0.199	(0.000)	0.195	(0.005)
		−24.4%	(1.3%)	−24.2%	(1.8%)	−29.0%	(0.1%)	−30.5%	(1.9%)
	R	0.792	(0.004)	0.791	(0.003)	0.524	(0.014)	0.606	(0.107)
		−1.9%	(0.5%)	−2.0%	(0.4%)	−35.1%	(1.8%)	−25.0%	(13.2%)
General	E	2.815	(0.011)	2.785	(0.015)	2.940	(0.005)	2.933	(0.029)
		13.3%	(0.5%)	12.1%	(0.6%)	18.3%	(0.2%)	18.0%	(1.1%)
	NMI	0.041	(0.003)	0.049	(0.004)	0.009	(0.001)	0.010	(0.007)
		−69.0%	(2.4%)	−62.7%	(3.1%)	−93.5%	(1.0%)	−92.2%	(5.2%)
	ρ	0.295	(0.008)	0.308	(0.013)	0.265	(0.000)	0.265	(0.003)
		−23.8%	(2.1%)	−20.7%	(3.3%)	−31.8%	(0.0%)	−31.6%	(0.9%)
	R	0.764	(0.004)	0.764	(0.003)	0.522	(0.013)	0.595	(0.095)
		−2.6%	(0.5%)	−2.7%	(0.4%)	−33.4%	(1.7%)	−24.2%	(12.1%)
Specific	E	3.063	(0.009)	3.043	(0.015)	3.152	(0.005)	3.148	(0.018)
		10.0%	(0.3%)	9.3%	(0.5%)	13.2%	(0.2%)	13.1%	(0.7%)
	NMI	0.033	(0.002)	0.038	(0.004)	0.010	(0.001)	0.011	(0.004)
		−70.3%	(2.2%)	−65.3%	(3.7%)	−90.7%	(0.8%)	−89.7%	(3.7%)
	ρ	0.295	(0.001)	0.295	(0.002)	0.297	(0.000)	0.297	(0.001)
		−7.7%	(0.4%)	−7.7%	(0.5%)	−6.9%	(0.1%)	−7.1%	(0.4%)
	R	0.772	(0.003)	0.771	(0.003)	0.521	(0.013)	0.601	(0.099)
		−1.8%	(0.4%)	−1.9%	(0.4%)	−33.7%	(1.7%)	−23.6%	(12.6%)
Term	E	6.676	(0.033)	6.634	(0.027)	7.758	(0.047)	7.567	(0.278)
		4.5%	(0.5%)	3.8%	(0.4%)	21.4%	(0.7%)	18.4%	(4.4%)
	NMI	0.241	(0.005)	0.247	(0.004)	0.078	(0.007)	0.107	(0.041)
		−15.4%	(1.8%)	−13.2%	(1.4%)	−72.6%	(2.5%)	−62.5%	(14.6%)
	ρ	0.112	(0.003)	0.107	(0.002)	0.108	(0.000)	0.107	(0.001)
		−7.4%	(2.4%)	−11.6%	(1.4%)	−10.7%	(0.2%)	−11.0%	(0.8%)
	R	0.873	(0.005)	0.871	(0.004)	0.532	(0.019)	0.635	(0.136)
		−1.2%	(0.6%)	−1.4%	(0.5%)	−39.8%	(2.1%)	−28.1%	(15.4%)

Table B.6. Text Clustering with Phrases: Medical Corpus: 11–14.

	Measures	11		12		13		14	
Combined	E	3.173	(0.014)	3.162	(0.014)	3.126	(0.010)	3.126	(0.008)
		9.6%	(0.5%)	9.3%	(0.5%)	8.0%	(0.3%)	8.0%	(0.3%)
	NMI	0.035	(0.004)	0.038	(0.004)	0.049	(0.003)	0.049	(0.002)
		−68.6%	(3.6%)	−65.8%	(3.3%)	−56.8%	(2.5%)	−56.9%	(2.1%)
	ρ	0.214	(0.004)	0.214	(0.005)	0.230	(0.006)	0.229	(0.007)
		−23.7%	(1.3%)	−23.8%	(1.7%)	−18.3%	(2.1%)	−18.5%	(2.3%)
	R	0.798	(0.001)	0.797	(0.004)	0.748	(0.008)	0.752	(0.009)
		−1.1%	(0.1%)	−1.3%	(0.5%)	−7.4%	(0.9%)	−6.9%	(1.1%)
General	E	2.814	(0.016)	2.795	(0.018)	2.753	(0.013)	2.747	(0.013)
		13.3%	(0.7%)	12.5%	(0.7%)	10.8%	(0.5%)	10.6%	(0.5%)
	NMI	0.041	(0.005)	0.047	(0.005)	0.058	(0.004)	0.060	(0.004)
		−69.0%	(3.4%)	−64.8%	(3.8%)	−56.0%	(2.7%)	−54.9%	(2.8%)
	ρ	0.309	(0.009)	0.313	(0.008)	0.321	(0.007)	0.322	(0.006)
		−20.2%	(2.3%)	−19.2%	(2.0%)	−17.4%	(1.7%)	−17.1%	(1.5%)
	R	0.771	(0.001)	0.769	(0.004)	0.728	(0.007)	0.732	(0.007)
		−1.8%	(0.1%)	−2.0%	(0.5%)	−7.3%	(0.9%)	−6.8%	(0.9%)
Specific	E	3.057	(0.012)	3.048	(0.011)	2.999	(0.006)	2.996	(0.008)
		9.8%	(0.4%)	9.5%	(0.4%)	7.7%	(0.2%)	7.6%	(0.3%)
	NMI	0.034	(0.003)	0.037	(0.003)	0.050	(0.002)	0.051	(0.002)
		−68.8%	(3.1%)	−66.6%	(2.7%)	−54.2%	(1.5%)	−53.5%	(1.9%)
	ρ	0.295	(0.001)	0.295	(0.001)	0.301	(0.004)	0.302	(0.005)
		−7.7%	(0.3%)	−7.7%	(0.2%)	−5.6%	(1.4%)	−5.4%	(1.5%)
	R	0.778	(0.001)	0.777	(0.004)	0.731	(0.007)	0.736	(0.008)
		−1.0%	(0.1%)	−1.2%	(0.5%)	−7.0%	(0.9%)	−6.5%	(1.0%)
Term	E	6.679	(0.022)	6.675	(0.021)	6.927	(0.052)	6.905	(0.064)
		4.5%	(0.3%)	4.5%	(0.3%)	8.4%	(0.8%)	8.0%	(1.0%)
	NMI	0.240	(0.003)	0.241	(0.003)	0.202	(0.008)	0.206	(0.010)
		−15.6%	(1.2%)	−15.4%	(1.1%)	−29.0%	(2.8%)	−27.8%	(3.5%)
	ρ	0.108	(0.002)	0.110	(0.002)	0.112	(0.002)	0.111	(0.002)
		−10.5%	(1.9%)	−9.1%	(1.3%)	−6.8%	(1.8%)	−7.7%	(1.5%)
	R	0.881	(0.001)	0.879	(0.006)	0.813	(0.010)	0.818	(0.011)
		−0.2%	(0.2%)	−0.5%	(0.6%)	−8.0%	(1.1%)	−7.4%	(1.3%)

Table B.7. Text Clustering with Phrases: Medical Corpus: 15–18.

	Measures	15		16		17		18	
Combined	E	3.105	(0.023)	3.089	(0.008)	3.116	(0.026)	3.090	(0.017)
		7.3%	(0.8%)	6.7%	(0.3%)	7.7%	(0.9%)	6.8%	(0.6%)
	NMI	0.054	(0.006)	0.059	(0.002)	0.051	(0.007)	0.058	(0.005)
		−51.7%	(5.6%)	−47.8%	(1.9%)	−54.4%	(6.3%)	−48.1%	(4.3%)
	ρ	0.227	(0.005)	0.219	(0.006)	0.231	(0.010)	0.237	(0.010)
		−19.3%	(2.0%)	−22.0%	(2.2%)	−17.7%	(3.4%)	−15.6%	(3.6%)
	R	0.797	(0.002)	0.795	(0.003)	0.769	(0.006)	0.772	(0.007)
		−1.3%	(0.3%)	−1.6%	(0.3%)	−4.8%	(0.8%)	−4.4%	(0.8%)
General	E	2.737	(0.022)	2.733	(0.007)	2.746	(0.029)	2.716	(0.020)
		10.2%	(0.9%)	10.0%	(0.3%)	10.5%	(1.2%)	9.3%	(0.8%)
	NMI	0.062	(0.006)	0.063	(0.002)	0.060	(0.008)	0.068	(0.006)
		−52.8%	(4.5%)	−52.0%	(1.5%)	−54.7%	(6.1%)	−48.3%	(4.3%)
	ρ	0.322	(0.011)	0.319	(0.007)	0.318	(0.014)	0.326	(0.004)
		−16.9%	(2.7%)	−17.7%	(1.8%)	−17.9%	(3.6%)	−15.9%	(1.1%)
	R	0.770	(0.002)	0.768	(0.002)	0.744	(0.006)	0.750	(0.005)
		−1.9%	(0.3%)	−2.1%	(0.2%)	−5.1%	(0.8%)	−4.5%	(0.6%)
Specific	E	3.008	(0.021)	2.999	(0.009)	2.982	(0.029)	2.958	(0.011)
		8.0%	(0.7%)	7.7%	(0.3%)	7.1%	(1.0%)	6.2%	(0.4%)
	NMI	0.048	(0.006)	0.050	(0.002)	0.055	(0.008)	0.062	(0.003)
		−56.4%	(5.2%)	−54.1%	(2.3%)	−49.8%	(7.4%)	−43.8%	(2.7%)
	ρ	0.295	(0.001)	0.294	(0.000)	0.301	(0.005)	0.300	(0.004)
		−7.6%	(0.4%)	−7.8%	(0.0%)	−5.7%	(1.6%)	−6.0%	(1.1%)
	R	0.776	(0.002)	0.774	(0.003)	0.751	(0.006)	0.755	(0.006)
		−1.3%	(0.3%)	−1.5%	(0.3%)	−4.4%	(0.7%)	−4.0%	(0.8%)
Term	E	6.599	(0.026)	6.592	(0.033)	6.781	(0.054)	6.755	(0.048)
		3.3%	(0.4%)	3.1%	(0.5%)	6.1%	(0.8%)	5.7%	(0.7%)
	NMI	0.253	(0.004)	0.254	(0.005)	0.225	(0.008)	0.229	(0.007)
		−11.2%	(1.4%)	−10.9%	(1.8%)	−21.1%	(2.9%)	−19.7%	(2.6%)
	ρ	0.116	(0.003)	0.111	(0.004)	0.110	(0.002)	0.108	(0.002)
		−3.6%	(2.9%)	−8.2%	(3.2%)	−9.0%	(1.4%)	−10.2%	(1.9%)
	R	0.876	(0.003)	0.874	(0.004)	0.840	(0.007)	0.843	(0.008)
		−0.8%	(0.4%)	−1.0%	(0.4%)	−4.9%	(0.8%)	−4.5%	(0.9%)

Table B.8. Text Clustering with Phrases: Newspaper Corpus: Tot, Rand, 1–2.

	Measures	Tot		Rand		1		2	
	E	2.320	(0.000)	2.310	(0.003)	1.373	(0.043)	1.227	(0.084)
		69.0%	(0.0%)	68.3%	(0.2%)	0.0%	(3.1%)	−10.6%	(6.1%)
	NMI	0.000	(0.000)	0.003	(0.001)	0.336	(0.015)	0.387	(0.030)
		−100.0%	(0.0%)	−99.0%	(0.3%)	0.0%	(4.6%)	15.4%	(8.9%)
ATot20.randPart1	ρ	0.212	(0.000)	0.233	(0.005)	0.660	(0.022)	0.704	(0.032)
		−67.9%	(0.0%)	−64.7%	(0.7%)	0.0%	(3.4%)	6.6%	(4.8%)
	R	0.200	(0.000)	0.020	(0.000)	0.802	(0.005)	0.814	(0.008)
		−75.1%	(0.0%)	−97.5%	(0.0%)	0.0%	(0.6%)	1.4%	(1.1%)

Table B.9. Text Clustering with Phrases: Newspaper Corpus: 3–6.

	Measures	3		4		5		6	
	E	2.258	(0.021)	2.254	(0.017)	2.241	(0.025)	2.246	(0.033)
		64.5%	(1.5%)	64.2%	(1.3%)	63.2%	(1.8%)	63.6%	(2.4%)
	NMI	0.024	(0.007)	0.025	(0.006)	0.030	(0.009)	0.028	(0.011)
		−93.0%	(2.1%)	−92.5%	(1.8%)	−91.0%	(2.6%)	−91.5%	(3.4%)
ATot20.randPart1	ρ	0.251	(0.010)	0.255	(0.005)	0.259	(0.013)	0.255	(0.018)
		−62.0%	(1.5%)	−61.3%	(0.8%)	−60.8%	(2.0%)	−61.4%	(2.7%)
	R	0.275	(0.016)	0.286	(0.015)	0.274	(0.019)	0.269	(0.027)
		−65.7%	(2.0%)	−64.3%	(1.8%)	−65.8%	(2.3%)	−66.5%	(3.4%)

Table B.10. Text Clustering with Phrases: Newspaper Corpus: 7–10.

	Measures	7		8		9		10	
	E	2.229	(0.020)	2.173	(0.024)	2.255	(0.021)	2.268	(0.011)
		62.4%	(1.4%)	58.3%	(1.7%)	64.2%	(1.5%)	65.2%	(0.8%)
	NMI	0.034	(0.007)	0.053	(0.008)	0.025	(0.007)	0.020	(0.003)
		−90.0%	(2.1%)	−84.1%	(2.5%)	−92.5%	(2.2%)	−93.9%	(1.0%)
ATot20.randPart1	ρ	0.271	(0.008)	0.305	(0.014)	0.253	(0.008)	0.248	(0.005)
		−58.9%	(1.1%)	−53.8%	(2.1%)	−61.7%	(1.2%)	−62.4%	(0.8%)
	R	0.333	(0.022)	0.426	(0.021)	0.280	(0.016)	0.285	(0.015)
		−58.5%	(2.7%)	−47.0%	(2.6%)	−65.1%	(1.9%)	−64.5%	(1.9%)

Table B.11. Text Clustering with Phrases: Newspaper Corpus: 11–14.

	Measures	11		12		13		14	
	E	1.499	(0.051)	1.511	(0.057)	1.374	(0.078)	1.401	(0.100)
		9.2%	(3.7%)	10.1%	(4.2%)	0.1%	(5.7%)	2.1%	(7.3%)
	NMI	0.291	(0.018)	0.287	(0.020)	0.335	(0.028)	0.326	(0.036)
		−13.4%	(5.4%)	−14.6%	(6.0%)	−0.1%	(8.2%)	−3.0%	(10.6%)
ATot20.randPart1	ρ	0.612	(0.023)	0.607	(0.028)	0.653	(0.034)	0.642	(0.040)
		−7.3%	(3.5%)	−8.1%	(4.2%)	−1.1%	(5.2%)	−2.8%	(6.0%)
	R	0.792	(0.005)	0.791	(0.005)	0.801	(0.008)	0.799	(0.010)
		−1.3%	(0.7%)	−1.4%	(0.6%)	−0.2%	(1.0%)	−0.4%	(1.3%)

Table B.12. Text Clustering with Phrases: Newspaper Corpus: 15–18.

	Measures	15		16		17		18	
	E	1.574	(0.090)	1.586	(0.063)	1.366	(0.063)	1.380	(0.061)
		14.7%	(6.6%)	15.5%	(4.6%)	−0.5%	(4.6%)	0.5%	(4.4%)
	NMI	0.264	(0.032)	0.260	(0.022)	0.338	(0.022)	0.333	(0.022)
		−21.3%	(9.5%)	−22.5%	(6.7%)	0.8%	(6.7%)	−0.8%	(6.4%)
ATot20.randPart1	ρ	0.576	(0.042)	0.571	(0.037)	0.651	(0.025)	0.642	(0.022)
		−12.7%	(6.4%)	−13.5%	(5.6%)	−1.4%	(3.7%)	−2.7%	(3.3%)
	R	0.786	(0.008)	0.786	(0.006)	0.801	(0.008)	0.802	(0.006)
		−2.1%	(1.0%)	−2.1%	(0.7%)	−0.2%	(1.0%)	−0.1%	(0.8%)