Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Juho Lehtonen

Clustering of Finnish scientific publications by discipline

Master's Thesis
Espoo, November 23, 2020

Supervisor: Professor Lauri Malmi
Advisor: Yrjö Leino Lic.Sc. (Tech.)

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Juho Lehtonen		
Title:	Clustering of Finnish scientific publications by discipline		
Date:	November 23, 2020	**Pages:**	55
Major:	Computer Science	**Code:**	SCI3042
Supervisor:	Professor Lauri Malmi		
Advisor:	Yrjö Leino Lic.Sc. (Tech.)		

Bibliometrics is a study of written scientific records such as books and articles in scientific journals. Its goal is to investigate how these outputs of research are connected and cite each other and what they can tell about science and its evolution. By studying scientific publications produced in Finland, the development of Finnish research can be monitored and, for example, compared to that of similar countries. The information thus formed helps, among other things, to make decisions that promote Finnish science and education. Designating scientific publications to the correct discipline is often a prerequisite for surveying that discipline. The currently commonly used classification of publications into disciplines is based on the journal level classification. All articles of a journal receive its classification. However, the journal is often classified into several disciplines, not all of which necessarily correspond to the topic covered by a single article. In addition, problems are caused by articles published in multidisciplinary journals, which now are left without a classification. A publication level classification that would more accurately identify the actual topic of a publication, is a subject of active research in bibliometrics. The goal is to create a classification based solely on attributes of the publication such as title, abstract and citations. One of widely studied methods is clustering. It is a commonly used machine learning method used to group similar observations into well separated clusters. Clustering is based only to the similarity determined by the selected features of the observations. The method does not need a pre-classified training set. In this thesis we look at the methods studied in bibliometrics to implement publication level classification. We describe the tested agglomerative hierarchical clustering method and the test data consisting of Finnish scientific publications. As a conclusion, we note that although some meaningful clusters seemed to emerge from the data, we did not achieve useful clustering of publications by discipline with this method.

Keywords:	clustering, bibliometrics
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Juho Lehtonen		
Työn nimi:			
Suomalaisten tieteellisten julkaisujen ryvästys tieteenaloittain			
Päiväys:	23. marraskuuta 2020	**Sivumäärä:**	55
Pääaine:	Tietojenkäsittelytiede	**Koodi:**	SCI3042
Valvoja:	Professori Lauri Malmi		
Ohjaaja:	Tekniikan lisensiaatti Yrjö Leino		

Bibliometriikka tutkii kirjoitettuja tieteellisiä julkaisuja kuten tutkimusartikkeleita ja -kirjoja. Sen tavoitteena on muun muassa selvittää, miten nämä tutkimuksen tuotokset liittyvät ja viittaavat toisiinsa sekä mitä ne voivat kertoa tieteestä ja sen kehityksestä. Suomessa tuotettuja tieteellisiä julkaisuja tutkimalla voidaan seurata suomalaisen tutkimuksen kehitystä ja esimerkiksi verrata sitä verrokkimaiden tutkimukseen. Näin muodostettu tieto auttaa muun muassa tekemään suomalaista tiedettä ja koulutusta edistäviä päätöksiä. Tieteellisten julkaisujen nimeäminen kuuluvaksi oikeaan tieteenalaan on usein esivaatimus kutakin tieteenalaa tarkasteltaessa. Nykyisin yleisesti käytössä oleva julkaisujen luokittelu tieteenaloihin perustuu julkaisukanavan eli tieteellisen aikakausilehden luokitteluun. Kaikki julkaisukanavan artikkelit saavat sen luokittelun. Julkaisukanava on kuitenkin usein luokiteltu useaan tieteenalaan, jotka kaikki eivät välttämättä vastaa yksittäisen artikkelin käsittelemää aihepiiriä. Lisäksi ongelmia luokitteluun tuottavat yleistieteelliset julkaisut, joissa julkaistut artikkelit jäävät vaille tieteenalaluokitusta. Julkaisukohtainen alaluokittelu, joka tarkemmin tunnistaisi julkaisujen varsinaisen aiheen, on bibliometriikan aktiivisen tutkimuksen kohteena. Tavoitteena on muodostaa pelkästään julkaisun ominaisuuksiin kuten otsikkoon, tiivistelmään ja viittauksiin perustuva luokittelu. Eräs paljon tutkituista menetelmistä on julkaisujen ryvästys. Ryvästys, eli klusterointi, on yleisesti käytetty koneoppimismenetelmä haluttaessa jakaa aineisto keskenään samankaltaisten havaintojen toisistaan mahdollisimman selkeästi erottuviin ryppäisiin. Ryvästys perustuu ainoastaan havaintojen valittujen ominaisuuksien määräämään samankaltaisuuteen. Menetelmä ei tarvitse valmiiksi luokiteltua opetusjoukkoa. Tässä työssä tarkastelemme bibliometriikassa tutkittuja menetelmiä julkaisukohtaisen luokittelun toteuttamiseksi. Kuvaamme kokeilemamme kokoavan hierarkkisen ryvästysmenetelmän sekä suomalaisia tieteellisiä julkaisuja sisältävän testiaineiston. Tämän työn tuloksena toteamme, että vaikka aineisto osittain vaikutti ryvästyvän mielekkäisiin julkaisuryhmiin, emme kokeillulla menetelmällä saavuttaneet käyttökelpoista julkaisujen ryhmittelyä tieteenaloittain.

Asiasanat:	ryvästys, klusterointi, bibliometriikka
Kieli:	englanti

Acknowledgements

I would like to thank my supervisors professor Samuel Kaski and professor Lauri Malmi for their kind help and patience. I am grateful to my instructor Lic.Sc. Yrjö Leino for his help and advice. Working at CSC – IT Center for Science has been a joy, thank you. Finally, I would like to thank my friends and family - and Piia.

Helsinki, November 23, 2020

Juho Lehtonen

Contents

1 Introduction **7**

2 Background **9**
2.1 Bibliometrics 9
2.1.1 Classification in bibliometrics 10
2.1.2 Bibliometrics in Finland 11
2.2 Clustering 12

3 Data and Methods **14**
3.1 Publication meta data 14
3.2 Feature extraction 15
3.2.1 Analysing textual data 15
3.2.2 Preprocessing 16
3.2.3 Lemmatisation 16
3.2.4 Vectorisation 16
3.3 Dimensionality reduction 17
3.3.1 Singular value decomposition 17
3.4 Model selection 18
3.4.1 Choosing the number of clusters 18
3.4.1.1 Internal validation of clustering results 19
3.4.1.2 Calinski-Harabasz criterion 19
3.4.1.3 Silhouette value 20
3.4.1.4 External validation of clustering results 20
3.4.1.5 Adjusted Rand index 21
3.5 Agglomerative clustering 21
3.5.1 Distance metric 21
3.5.2 Linkage methods 22
3.5.2.1 Single linkage 22
3.5.2.2 Complete linkage 22
3.5.2.3 Average linkage 22
3.5.2.4 Ward's method 22

3.5.3 Algorithmic complexity 23

4 Implementation **24**
4.1 Preprocessing 24
4.2 Tokenizing 25
4.3 Lemmatising 25
4.4 Removing stop words 26
4.5 Vectorizing 26
4.6 Clustering 26

5 Results **30**
5.1 Manually annotated data 30
5.2 Finnish publications from years 2000-2001 33

6 Discussion **39**

A Manually annotated data set **47**

B Top terms **49**

C Articles **50**

Chapter 1

Introduction

This thesis handles the problem of clustering Finnish scientific publications by their meta data. The goal is to cluster them as well as possible by their scientific discipline. We want to find a clustering for the publications such that it matches as well as possible an imaginary average manual classification done by human experts.

Classifying publications (such as research articles) by discipline enables different types of bibliometric analyses. Such classifications can be used in summarizing research output of a university or a country. The sum of all scientific publication of an individual country, for example, can be interesting and useful information that can tell something about research and its level in that country.

We use Finnish scientific publications from Web of Science (WoS) database as our source data. Web of Science database gets approximately 10000 publications annually that can be assigned as a result of Finnish research, that is, at least one author is affiliated to a Finnish research organisation (see [2] for definition). It should be noted that not all Finnish research can be found from WoS database. Especially humanities and social sciences are under-represented, so it isn't a comprehensive image of Finnish research. Due to the amount of the scientific publications, manual classification is not applicable. It is also impossible for any one person to master all fields of science to manage such a task. Our goal is a method that the analyst herself can apply.

For clustering the publications, we will use meta data that the authors of the publication will provide in any case with their work such as the title, the abstract and the keywords. Clustering will not produce any meaningful labels for the clusters. They must be labelled appropriately by other means.

The "wellness" of a clustering measured by how it fits to scientific fields is a tricky issue for at least couple of reasons. First, there is no general consensus about what is the correct partition of all science to its different branches, or even what are the different fields of science. It might depend on specific need or individual opinion where the line between two related discipline lies. Second, science is evolving all the time. What was perhaps at some point in history seen simply as chemistry is today commonly divided at least to organic and inorganic chemistry. Third, the definition of scientific disciplines depends also how closely we look into a topic. Sometimes chemistry is sufficient a classification, for example, for a publication and sometimes we need to label it more accurately as organic chemistry. Fourth, there is a considerable amount of research that is intrinsically cross-disciplinary. Bioinformatics, for example, relies heavily on biology and computer science at least. So, it depends on the purpose whether a publication should be considered belonging to biology, computer science, bioinformatics or to all of them.

In Chapter 2, we will shortly present the general concepts of bibliometrics and some efforts in the field in Finland. We will also present the main focus of this thesis, clustering. Next in Chapter 3, we describe the data and present the methods we will use to analyse it. In Chapter 4, we go through the steps of the analysis in practice describing all the choices we made. After that in Chapter 5, we present the results and finally summarize all in Chapter 6.

Chapter 2

Background

In this chapter, we briefly introduce clustering and how it is positioned in the larger field of machine learning. But first, we describe what bibliometrics is.

2.1 Bibliometrics

Bibliometrics is a study of written scientific records. The records may be books, articles, letters in scientific journals, conference papers and so on. Bibliometrics studies how these products of research are communicated, how are they related to each other, what kind of properties they have and what can be learned about the science in general by analysing them.

Bibliometrics seeks to answer questions like “How many publications has an author authored?”, “How many citations an author has”, “What are the cited publications of a scientific document?”. It also studies a bit more broader questions like “How many publications there has been published on a discipline X a year?”, “Which research area does this publication belong to?”, “What other publications belong to this research area?”, “When has this research area emerged?”, “What are the most important related research areas of this discipline?” and so on.[32][13]

Classifying things is often the first thing we do when we observe the world. On the other hand, counting the frequencies of objects and comparing these numbers often helps to put things in perspective. One of the earliest studies that is generally considered bibliometrics was Cole’s and Eales’ analysis to the anatomy literature in 1917 [15]. In their study they researched the anatomy literature from 1543 through 1860 with the intention to graph the growth of the number of documents over the three centuries, to present “the performance” of each European country, to observe the most popular topics among scholars from time to time, and to compare the advancement and devolution stages of the research with different societal events [10]. As the number of scientific publications has enormously increased the need to automatically analyse them has become apparent. The basic analysis on top of which more detailed studies can be built on is classifying each publication to research areas and disciplines.

2.1.1 Classification in bibliometrics

Currently, the most popular systems to classify the publications into research areas are Clarivate's (formerly Thomson Reuters) Web of Science[1] and Elsevier's Scopus[2] classification systems. These classification systems classify journals into one or more research areas [40]. Publications are then classified to research areas based on in which journal they were published. WoS uses approximately 250 *subject categories* in its classification. Each journal can belong to one or up to six categories. The categories have been created latest on early 1960's by manually classifying journals. New journals were added one at the time after visual inspection of citation information. New categories were added when old ones grew [33]. As for Scopus, according to Wang and Waltman [41], "there seems to be no information at all on the construction of its classification system". We will use the data from WoS database in this work.

Also an independent journal level classification system has been developed [9]. Journal level classification systems have known limitations. They are, for example, unable to meaningfully classify publications published in multi-discipline journals. Also some discipline specific classification systems exists such as Association for Computing Machinery's Computing Classification System for computer science discipline [3], the U.S. National Library of Medicine Classification [6] for medical resources and American Physical Society's Physics Subject Headings [4] for physics publications for example. An alternative classification system is publication level classification where each publication is classified based on some information extracted directly from the publication self. Commonly used information is the title, the abstract, keywords attached by the authors or the publisher and citations to other publications. Shu et al. have compared journal and paper level classification approaches and found that publication level classification could provide better classification [35].

Bibliometric research uses mainly three types of methods for journal or

[1] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[2] https://www.scopus.com/

publication level classification; citation based, text analysis based or combination of the two [22]. Citation based methods study citations of publications and produce networks where publications are nodes connected by citations as edges. In the rare case of publication having a direct quote including a citation, the citation is not counted unless it is also a citation of the publication itself. Connection between two publications can be formed by a direct citation, bibliometric coupling where publications are connected when a third publication cites them both or co-citations where publications are connected if they cite the same third publication.

Text analysis based methods can examine, for example, the title, the abstract or the whole text content of the publication itself and classify the publications or journals by the topic model created [11]. Hybrid models combine both approaches. Authors and their affiliations are not used in this work because we cannot uniquely identify them and we can not assume their field of science that is our interest here. Classifying publications into scientific disciplines is an open problem and a subject of active research. Currently, there is no known general, good enough practical solution.

2.1.2 Bibliometrics in Finland

Ministry of Education and Culture of Finland provides yearly updated bibliometric analyses of Finnish research activities based on both Web Of Science citation index and Elsevier's Scopus database [5, Vipunen service]. The corresponding source system classification for a field of science is used and aggregated to match the Statistics Finland classification [2]. CSC - IT Center for Science is responsible of the actual technical implementation of the service.

One of the earlier and comprehensive bibliometric research of Finnish science is a report by Persson et al. [31] which mapped the situation and development of Finnish science 1981-1998. This, however, is a report which concentrates on bibliometric analysis based on the map of science provided by WoS subject categories. But if we want to explore how the Finnish scientific disciplines themselves have evolved over time, these pre-defined subject categories are quite a rigid framework. For that reason, we want to experiment creating an alternative mapping of science by clustering. So, as opposed to bibliometric analyses seeking to understand the state of a scientific discipline as defined by some existing definition, we want to experiment and inspect how to define scientific disciplines to be used in bibliometric analyses.

Suominen and Toivanen [38] used unsupervised learning-based topic modelling to create a map of science for Finnish publications from 1995-2011. They evaluated it by comparing the results with WoS based classification and concluded that superiority of the method depends on the purpose of the analysis. Traditional manually created classifications are relevant for information retrieval while machine learning methods can reveal new emerging areas of science. Compared to our work here, their analysis method topic modelling differs from hierarchical clustering albeit both are unsupervised learning methods.

Our research question here is: *"How to automatically cluster Finnish scientific publications and how does that clustering compare to existing fields of science classification by WoS?"* We will use hierarchical clustering on features derived from titles, abstracts and keywords in publication meta data.

2.2 Clustering

The methods discussed in this work belong to the field of machine learning. The field has its roots in statistics and engineering and is itself part of artificial intelligence, a sub-field of computer science. The methods in machine learning can be divided to *supervised*, *unsupervised* and *reinforced learning* [8].

Commonly for all methods we define X as the sample data and Y as label indicating which class our data sample X belongs. We also have to choose the model $f()$ for learning from the data. Then we can state our learning problem as a function $Y = f(w \cdot X)$, where w is a weight vector, which would give a prediction of the class Y of the sample data X. Assuming we have enough of samples X we then teach our model with training data. That is, we solve the weights w using the loss function such that it optimizes the difference between the true and the predicted class labels Y.

Supervised learning methods include classification and continuous estimation (regression) where the class of the training samples, or the target values in case of regression, Y is known. Example of a classification problem is optical character recognition where the system is taught with example of characters along with their correct labels.

For unsupervised learning the correct answer or the label Y is not known. Instead a model is applied such that it finds regularities in the input data X. Example of unsupervised learning problem is finding anomalies that do not fit in the group, such as analysing log files of a computer system to find a possible intruder.

Unsupervised learning methods include clustering, dimensionality reduction and topic modelling for instance. By clustering we try to distinguish patterns in the data and discriminate unrelated objects into separate clusters and aggregate related objects into same cluster.

In reinforced learning we want the system to learn a sequence of actions leading to desired outcome. For example, we may want the system to learn to win a game. In that case individual actions are not important but the end result as there are many ways to win a game. So, the system repeatedly tries different combinations of actions while it receives the result of its combined actions.

Chapter 3

Data and Methods

In this chapter we present the data and methods used in the clustering. We follow the logical order in which the methods are applied on the data. We first describe the raw data and pre-processing it. Then we explain the feature extraction transforming irregular length symbolic data (text) into numerical vector representation. Next to follow is reducing the dimensionality of the data, selecting the model used to learn from the data and finally the selected clustering method.

3.1 Publication meta data

The data consists of 21155 records of Clarivate's Web of Science publication meta data from years 2000-2001. Each record describes some basic information about an article published in a scientific journal. The data contains only publications with at least one author with an affiliation to a Finnish research organisation as recorded in the publication. These publications were published in total in 3518 different scientific journals. An example of a shortened record:

```
Lehti: ACTA OPHTHALMOLOGICA SCANDINAVICA
ISSN: 1395-3907
Ala: OPHTHALMOLOGY
Ilmestymisvuosi:  1999
Otsikko: Assessment of diabetic retinopathy using two-field 60
degrees fundus photography. A comparison between[...]
Abstrakti
Purpose: To assess the severity of diabetic retinopathy and
maculopathy bycomparing[...]
Avainsana (KeywordPlus):  OPHTHALMOSCOPY
```

```
Avainsana (KeywordPlus):  KAPPA
[...]
Avainsana (tekijät):  diabetic retinopathy
Avainsana (tekijät):  diabetic maculopathy
[...]
Lähde: 0010603696 /  *DIAB CONTR COMPL /  NEW ENGL J MED /  977
/  329 /  1993
Lähde: 0034118371 /  *DIAB CONTR COMPL /  ARCH OPHTHALMOL-CHI /
1344 /  105 /  1987
Lähde: 0075276068 /  *DRS RES GROUP /  OPHTHALMOLOGY /  82 /  85
/  1978
```

Each meta data record contains a single instance of the title ("`Lehti`"), ISSN (International Standard Serial Number), the field of science ("`Ala`") and publication year ("`Ilmestymisvuosi`") of the journal, the title ("`Otsikko`") and the abstract ("`Abstrakti`") of the article. Additionally each meta data record can contain multiple instances of keywords inferred by the publisher programmatically ("`Avainsana (KeywordPlus)`"), keywords produced by the authors themselves ("`Avainsana (tekijät)`") and the cited references ("`Lähde`"). The data contains incomplete, erroneous and multiplicated records.

3.2 Feature extraction

The clustering algorithms don't understand text documents but require numerical input. To enable the handling of the textual data by the clustering algorithms we have to transform it into a numerical form. We usually describe the data as a matrix where each *data sample*, or observation, a single publication in our case, is one row in the matrix. Each data sample consists of *features*, that is numerical values representing some aspect of the sample. So it follows that the rows of the matrix are feature vectors of the data samples and the columns are the individual features. Feature extraction is the process used to transform text documents into feature vectors (i.e. data samples). The number of columns in the data matrix corresponds to the order of the feature space of the task.

3.2.1 Analysing textual data

The descriptions are natural English language appended with the citation references. When analysing this kind of textual data the often used methods involve some kind of counting. We count, for example, to find the most used words in a document, which words appear together and so on. Next we will describe the methods used in this work.

3.2.2 Preprocessing

The preprocessing for text analysis usually also includes the removing of *the stop words* from data. Stop words are the most frequent words in the data like: *"the, of, and, where etc."* Because these words are present in any text they probably don't tell much about the topics these records concern as presented by Luhn [28].

3.2.3 Lemmatisation

After removing stop words the next step is to unify the different written forms a term. We might have, for example "visual" and "visually" or "dog" and "dogs" and we want them as "visual" and "dog" only. This is desirable to reduce the redundant repetition of the data and also to reduce the dimensionality of the feature space [36][19]. There are two possible options to achieve this.

Lemmatisation means replacing each inflectional form of a word with its nominative (i.e. dictionary) form, or lemma. The problem is the ambiguousness of many natural words. To achieve lemmatisation, many tactics from simple dictionary look-ups to rule-based systems, to sophisticated algorithms to infer the role of the word in the sentence as well as using the larger context are employed. The stop words should be in their nominative form so they should be removed after lemmatisation.

Stemming means stripping the word of its termination such that only the stem of the word is retained. No context is used and only the word itself is inspected. It is much simpler procedure compared to lemmatisation. Lemmatisation results in better precision, or true negative rate but poorer recall, or true positive rate, compared to stemming. [29]

3.2.4 Vectorisation

After lemmatising the terms we count the occurrences of each term in each document. This is called *the bag of words* representation of the textual data because for each word in the corpus, we only count it ignoring all its positional information in relation to other words. So each document is represented by a vector of its term frequencies. This vector is called *the feature vector* of the document. The resulting term occurrence frequencies are normalized to decrease the importance of the tokens that occur in the majority of documents. Usually these are common terms not specific to the topic of the document. These normalized occurrence frequencies are called term frequency inverse document frequencies, TF-IDF, and they form the features of a document. The size of the feature space is determined by the number of counted unique terms in all documents of the collection. TF-IDF is defined as follows:

$$TF - IDF(t, d) = tf(t, d) \times log\frac{1 + n}{1 + df(t)} + 1, \tag{3.1}$$

where $tf(t, d)$ is the term frequency, count of a term t in a document d, n is the number of documents in the data set and $df(t)$ the number of documents containing term t [27][23].

Here the size of the feature space is in the order of the the size of the English language corpus i.e. in tens of thousands. The document feature matrix is very sparse, a single document only having handful of terms of the whole corpus.

3.3 Dimensionality reduction

The resulting publication-feature matrix has the dimensionality of $m \times n$ where m is the number of records and n is the number of features and very sparse. In our case the record-feature matrix has the dimensionality of approximately 21000×50000. So the dimensionality of our problem, number of features, is very high compared to the number of samples. This means that all samples are very sparsely and very far away from each other in the feature space. The higher the number of dimensions the more dissimilar each observation seems from every other observation. This is known as *the curse of dimensionality* [39]. To tackle this, and also to fasten the computation of the clustering we have to reduce the dimensionality of the feature space.

3.3.1 Singular value decomposition

In text analysis setting the logical reasoning for dimensionality reduction is known as Latent semantic analysis (LSA)[17]. LSA is based on a linear algebra tool called singular value decomposition (SVD). SVD is a common method to reduce the dimensionality of data. SVD for a matrix M is defined as:

$$M = U\Sigma V^T \tag{3.2}$$

Here U is unitary matrix, Σ is the diagonal matrix of singular values σ_i of M and V consists of orthonormal set of singular vectors of M.

SVD transforms data linearly into "a new feature space" in which the features are not correlated. In the decomposition the new features are ordered in decreasing order of variance so the least significant ones can be omitted. Another commonly used method, principal component analysis (PCA), achieves the same. Whereas PCA requires the data matrix to be square and have eigendecomposition, SVD does not, but achieves the same goal. Because all matrices do have SVD but not necessarily eigendecomposition, SVD is more general method.

3.4 Model selection

Generally all machine learning problems are ill-posed in the sense that a unique solution for the problem can't be found unless some assumptions, or *inductive bias*, are introduced. This begins with selecting the learning algorithm and might also include some hyperparameters of selected algorithm.

Here we will use unsupervised learning to shape the mapping of scientific disciplines because we want to learn the possible intrinsic structure of scientific knowledge. There are many different clustering algorithms from which to choose. Some often used common algorithms are k-means, hierarchical clustering, density based scan clustering and Gaussian clustering. We will use hierarchical clustering because it produces hierarchical cluster structure that is naturally expected of fields and sub fields of science. In hierarchical clustering there are yet different parameters to choose. Hierarchical clustering with single, average and complete linkages and Ward's method applied to search query result clustering were studied by Korenius et al. [25].

3.4.1 Choosing the number of clusters

The number of clusters k that should result from the clustering is not known beforehand but is a required parameter of the clustering algorithm. Hierarchical clustering will always return k clusters regardless if they are meaningful or not. To make a decision about the number of clusters we can 1) inspect the data in two dimensions using some dimensionality reduction method such as PCA and then plot it, 2) set limits for the inter-cluster distances at each step of merging clusters, that is, finding the largest gap between the dendrogram levels or 3) measuring the overall compactness of clusters and their separation from each other over the number of clusters k with suitable evaluation criterion such as average silhouette value [8][14][34]. The first method is assumed to give uninformative view of the data. We will use the third method. We expect to observe a maximum in the criterion value indicating a “natural” cluster structure in the data. Currently the number of manually curated WoS subject categories is 254[1] so we position our initial expectations near that.

3.4.1.1 Internal validation of clustering results

Liu et al. have reviewed 11 commonly used internal clustering validation indices [26]. These are used to decide the correct number of clusters. We will use two of those; Calinski-Harabasz criterion [14] and Silhouette value [34]. We choose these because they were familiar for us and have different weak spots as noted by Liu et al. and thus could complement each other.

3.4.1.2 Calinski-Harabasz criterion

Calinski-Harabasz criterion is defined as

$$CH = \frac{SS_B}{SS_W}\frac{N-k}{k-1}, \tag{3.3}$$

where SS_B is *the between-group sum of squares* that gives the overall variance between clusters

$$SS_B = \sum_{i=1}^{k} n_i ||m_i - m||^2, \tag{3.4}$$

and SS_W is *the within-group sum of squares* that gives the overall variance within clusters

$$SS_W = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - m_i||^2. \tag{3.5}$$

In equations above k is the number of clusters, N is the number of observations, n_i is the number of observations in cluster i, m_i is the centroid of cluster i, m is the overall mean of sample data, x is the observation, C_i is the ith cluster and $|| \cdot ||$ is the Euclidean distance between the two vectors. The larger the Calinski-Harabasz criterion, the better the cluster structure.

[1] https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/

3.4.1.3 Silhouette value

Averaging over all observations in all clusters gives the average silhouette value for the whole clustering:

$$S = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{x\in C_i} s(x). \tag{3.6}$$

Here the silhouette value $s(x)$ of a single clustered observation x is defined as follows.

$$s(x) = \frac{b(x) - a(x)}{max(b(x), a(x))}, \tag{3.7}$$

where $a(x)$ is the average dissimilarity of the observation x to all other observations in its own cluster C_i:

$$a(x) = \frac{1}{n_i - 1}\sum_{y\in C_i, y\neq x} d(x, y) \tag{3.8}$$

and $b(x)$ is the average dissimilarity of the observation x to all observations y in the next closest cluster C_j:

$$b(x) = min_{j,j\neq i}[\frac{1}{n_j}\sum_{y\in C_j} d(x, y)] \tag{3.9}$$

In equations above k is the number of clusters, n_i is the number of observations in cluster i, x is the observation, C_i is the ith cluster and $d(x, y)$ is some dissimilarity measure on ratio scale (e.g. Euclidean distance) between the two vectors. Silhouette values are bounded to range $[-1, 1]$. Values close to 1 mean that the observation x is very well clustered in its cluster, value 0 means that the observation could as well be clustered in the nearest neighbour cluster and values close to -1 mean that the observation is almost certainly in the wrong cluster.

3.4.1.4 External validation of clustering results

To support the choice of internal validation method for choosing the number of clusters, we will use the external validation with a manually annotated data set. Goal will be to achieve evaluation measurements based on a quite clearly separated set of publications. We will create a manually annotated validation set providing the ground truth needed for calculating the adjusted Rand index. The validation set will consist of publications from three different disciplines. We will experiment which of the internal validation methods, Silhouette value or Calinski-Harabasz criterion, will indicate more clearly the known, correct number of clusters.

3.4.1.5 Adjusted Rand index

Adjusted Rand index measures the similarity of the clustering result with our defined ground truth clustering. We will compare the internal validation result over the clusterings with different number of clusters with Adjusted Rand index (ARI) [21] to see how well our expected number of clusters 3 stands out. Adjusted Rand index is defined as follows.

$$ARI = \frac{RI - E[RI]}{max_{perm}(RI) - E[RI]}, \tag{3.10}$$

where RI is the "raw" Rand index, $E[RI]$ its expected value and $max_{perm}(RI)$ is its maximum value over all possible clustering permutations. The Rand index is defined as:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{3.11}$$

where TP is the number of true positives, TN true negatives, FP false positives and FN false negatives. ARI values are bounded to range $[-1, 1]$ where the perfect clustering would be 1.

3.5 Agglomerative clustering

In this work we use agglomerative hierarchical clustering with Ward's distance metric [42]. This method was chosen because it was familiar to us, the data naturally has hierarchical structure of fields and sub fields of science that goes from general disciplines to more and more specialized research topics and is known to perform well enough in simple settings. Agglomerative clustering starts with each document as a cluster of its own and then merges pairs of clusters together as the clustering process goes on.

3.5.1 Distance metric

Our data is high-dimensional so choosing distance measure is important. Boyack et al. have compared clustering real world size corpus of 2.5 million publications from MEDLINE with nine different similarity metrics [12]. For example Manhattan distance works well with sparse data.

$$||a - b||_1 = \sum_i |a_i - b_i| \tag{3.12}$$

Euclidean distance is the familiar distance used in everyday life:

$$||a - b||_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{3.13}$$

Because Ward's method assumes Euclidean distance, we use it.

3.5.2 Linkage methods

Linkage method affects how clusters are formed and how the distance metric is applied between two clusters. In following $d(a, b)$ denotes the distance between observations a and b.

3.5.2.1 Single linkage

Single linkage defines that the distance between two clusters A and B is measured as distance between two closest items of the clusters.

$$d(A, B) = min\{d(a, b) : a \in A, b \in B\} \tag{3.14}$$

3.5.2.2 Complete linkage

Complete linkage is defined as the distance between the most distant items of clusters A and B.

$$d(A, B) = max\{d(a, b) : a \in A, b \in B\} \tag{3.15}$$

3.5.2.3 Average linkage

Average linkage is defined as the average of the all pairwise distances between the items of cluster A and B.

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{3.16}$$

3.5.2.4 Ward's method

We are using Ward's method because it usually creates compact even-sized clusters [37]. Ward's method minimizes the total within-cluster variance for merging the next possible clusters A and B:

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}(\bar{\mathbf{a}} - \bar{\mathbf{b}})'(\bar{\mathbf{a}} - \bar{\mathbf{b}}) \tag{3.17}$$

where $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are the centroids of cluster A and B respectively and n_A and n_B are corresponding number of members in clusters A and B.

3.5.3 Algorithmic complexity

Time complexity of the used agglomerative clustering with Ward's method is $O(N^3)$ and space complexity is $O(N^2)$, where N is the number of publications [43]. These are quite demanding requirements when the number of publications increase. In this case it is acceptable because there is no need for real time updates of results, rather we experiment with exploratory data analysis.

Chapter 4

Implementation

We first describe briefly the used programming language and environment and then how the methods are implemented to achieve the clustering.

We implemented the whole workflow with Python 3 programming language. The core clustering and performance evaluation modules were imported from Python's `scikit-learn` package [30]. `scikit-learn` is a widely used open source package for machine learning, providing functions for classification, regression, clustering and pre-processing amongst other things. It is openly developed by international community of developers. We wrote functions for pre-processing, feature extraction, clustering and plotting around these modules. The top level workflow management was written using Python's `doit` workflow management package. In total this work amounted roughly 1500 lines of functioning Python code, approximately half of which is our contribution and the rest from open source documentation. All code is publicly available[1]. Due to licensing agreements between Ministry of Education and Culture and Clarivate the data to reproduce the results is not publicly available. The overall workflow is depicted in Figure 4.1. The analysis was run on laptop environment with Fedora 28 Linux 5.0.16-100.fc28.x86_64, Intel Core i7-4800MQ CPU @ 2.70GHz, 16 GB of RAM.

4.1 Preprocessing

The data was first read from raw files obtained from publishers. We choose to omit some meta data fields from the analysis (cf. section 3.1). Fields *journal*, *issn* and *the field of science* were omitted because they exhibit the existing classification whereas we wanted an alternative one. The field *year* is irrelevant here as our analysis doesn't include temporal aspect. The field

[1] https://github.com/juholehtonen/clustering/releases/tag/rc-1.3

references was also left out because the references would have needed different treatment and more theoretical research on how to be used as features for clustering.

We choose to keep *title*, *abstract* and *keyword* fields. All the title, abstract and author provided keywords are original data of the publication without any external interpretation. We also choose to keep the automatic publisher inferred keywords ("`Avainsana (KeywordPlus)`") because these can be seen as concentration of publication data by some text analysis algorithm. The algorithm is unknown here but we assume that the aim of the these keywords is also to describe the publication as well as possible. Terms in keyword fields were concatenated into combined terms: "allergic_contact_dermatitis". This is related to the counting of words in a meta data and will be explained soon. The kept meta data fields were read into a dictionary and data saved in intermediate files.

4.2 Tokenizing

The meta data in chosen fields was then tokenized with whitespace and punctuation other than periods separating the tokens (words). We get for example "*Pluto is a smart dog*" → "*Pluto*", "*is*", "*a*", "*smart*" and "*dog*" (the *bag of words* representation). We made an exception in tokenizing the keywords as explained above. Alternatively we could form n-grams out of the text to preserve the meaning of combined terms. For example when tokenizing "*allergic contact dermatitis*", instead of "*allergic*", "*contact*" and "*dermatitis*" we would get "*allergic contact*" and "*contact dermatitis*". Here single words are called 1-grams, word pairs are called 2-grams and so on. Counting the n-grams would expand the feature space n-fold though and we wanted to avoid the computational cost.

4.3 Lemmatising

After tokenizing the words we lemmatised them with NLTK's[2] WordNet lemmatiser. The lemmatiser uses WordNet lexical database developed and maintained by Princeton University [1]. It is a large general purpose English[3] language database with over 155000 words and over 175000 synonym groups. We choose to remove all plain numbers from data. Another option is to replace numbers with dummy '#NUMBER'. That could help separate natural

[2] https://www.nltk.org/, a Python package for natural language processing

[3] Other languages are supported as well.

sciences from humanities. Because we did not have proper knowledge on the issue, removal was chosen. We assumed that to be a more neutral way to treat them. If the lemmatiser did not find the word in WordNet Lemmatiser database, the word was returned unchanged.

4.4 Removing stop words

Stop words were then removed after lemmatisation. We used Python's `NLTK` module's English stop words combined with `scikit-learn's` stop words and our own short stop word list consisting standard publisher notices: (*"rights"*, *"reserved"*, *"science"*, *"elsevier"*, *"2000"*). The stop word list included 383 words.

4.5 Vectorizing

After removing stop words the term frequencies were counted and normalized with inverse document frequencies (*tf-idf* normalization). In this point we have the data vectorized.

We set the minimum document frequency, under which the terms having their *tf-idf* value are ignored, to $min_d f = 2$. Thus, terms that have less than two occurrences are ignored. The maximum document frequency is set to $max_d f = 0.1$. So, terms occurring in more than 10 % of the documents are also ignored.

4.6 Clustering

When clustering with Ward's hierarchical clustering the interesting parameter values are the number of the clusters, the distance measure, the cluster connectivity measure and the number of the components or the amount of dimensionality reduction.

To decide the number of clusters we experimented with the manually annotated data set described in section 3.4.1.4. The validation set consists of Finnish publications from year 2000 from three different fields of science as categorized by WoS. Two of the fields are closely related sub fields of computer science: *computer science: information systems* (CS-IS) and *computer science: artificial intelligence* (CS-AI), while the third one, *clinical neurology*, is more distant from those. There are 116, 122, 250 publications labelled as belonging to fields CS-IS, CS-AI and clinical neurology respectively, and

31 publications labelled as belonging to both CS-IS and CS-AI, totalling 519 publications in our data.

Publications were inspected by title, abstract, keywords, journal and publisher assigned disciplines of the journal. Publications with critically missing data, unclear discipline assignment and more like very application specific topics (e.g. a publication describing the development of a virtual community and a community game platform) were excluded from validation set. We then checked with layman's reasoning if the WoS labelled discipline seemed plausible. Because of the journal based classification, most publications had more than one assigned discipline. Only the previously named three discipline labels were retained because we wanted to test if and how well our clustering method would cluster the samples compared to these three groups. We asked an another opinion for 37 the most ambiguous publications that could have been in more than one of our categories. The second opinion was given by an academically educated person who was not an expert of all the fields in the data. Decision for these was made on consensus or by draw if needed. For manually curated validation set with discipline assignment and justifications for possible exclusion see appendix A. In total 29 publications were manually reclassified by either choosing only one of two used classes or by assigning the publication to more fitting category by our judgement. After excluding bad data and reclassifications we had a data set of 134, 127 and 194 publications in CS-IS, CS-AI and clinical neurology, respectively, totalling 455 publications.

The basic problem is that fields of science can not be defined so that they clearly differ from each other. Where one discipline ends the other has already started like metallurgy and material science. Likewise, there are lots of publications that handle topics belonging to more than one discipline, e.g. this thesis discusses clustering and bibliometrics. So when annotating publications, deciding if a publication belonged to a discipline or not felt often quite difficult. Often the separation between disciplines felt quite arbitrary. For example an article describing using wavelet transformation for coding noisy images was decided to belong to CS information systems whereas an article describing wavelet based corner detection using SVD was decided to belong to CS artificial intelligence. For CS artificial intelligence we mostly selected publications which mentioned some dimensionality reduction or machine learning related term or concept. CS information systems ended up being quite like some "others" or "the rest" dump class. It would have publications such as "A reference model for conceptualising the convergence of telecommunications and datacommunications service platforms", "Developing a distributed meeting service to support mobile meeting participants", "On voice quality of IP voice over GPRS". Decision to assign publications to one discipline only was based on using hierarchical and thus hard clustering. Alternatively, soft, or probabilistic, clustering would allow a publication to be clustered to multiple clusters with varying probabilities. For clustering the whole data for years 2000-2001 we chose the number of clusters based on silhouette values and Calinski-Harabasz criterion.

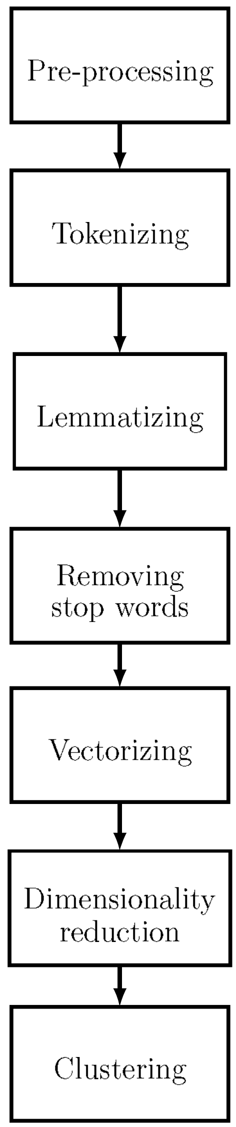

Figure 4.1: The workflow of the clustering.

Chapter 5

Results

We first calculated internal and external validation metrics for the manually annotated subset of data to see how well our metrics revealed our defined ground truth clustering. Based on this we tried to evaluate the most suitable internal validation metric to be used with the complete data set. This validation metric would suggest us the optimal number of clusters k to be used in the clustering.

5.1 Manually annotated data

We evaluated our internal clustering validation metrics by creating a smaller subset of publications with manually checked field of science (as described in Section 4.6). This subset consisted of 455 publications from three different fields: *Computer science: Information systems*, 134 publications, *Computer science: Artificial intelligence*, 127 publications and *Clinical neurology*, 194 publications. The term frequency cut-off values $min_d f = 2$ (number of occurrences) and $max_d f = 0.5$ (portion of documents with the term) resulted in 4455 ignored terms and left the feature space of 3177 terms. Terms that were filtered out as too frequent or too rare were for example: "haemoglobin_a", "jacobian_matrix", "parthenogenetic", "chelation_by_saccharide_molecules", "leg_blood_supply", "computational_fluid_dynamics", "pigment", "hdtv", "nicotinic_receptor".
(Note: keywords were combined with '_' as described in Section 4.1.)

The dimensionality of data was then reduced with LSA from 3177 features (corresponding to the terms) to 800 components which resulted 100 % of the explained variance of the data to be retained. The data was then agglomeratively clustered with Ward's method for number of cluster values $k = [2, 12]$. The Figure 5.1 shows the internal clustering validation results for these clusterings. We see that Calinski-Harabasz index reaches its maximum

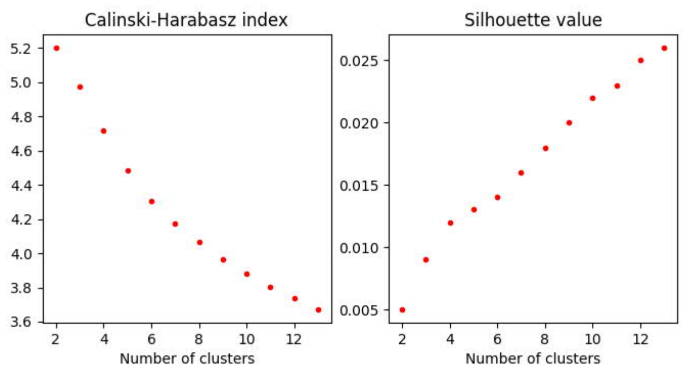

Figure 5.1: Calinski-Harabasz index and average silhouette values for the manually annotated set of 455 publications clustered with hierarchical clustering, LSA with 800 components. The higher value means more compact and better clustering in both graphs.

value with the number of clusters at two and that silhouette values increases with the number of clusters. Neither of these indices suggests three clusters as the optimal clustering solution. So according these specific metrics the data doesn't support our initial idea of three "natural" clusters as we meant it. We repeated the clustering in case of ambiguousness in the algorithm but that didn't affect the clustering result. Probably the low amount of samples compared to still quite high dimensional feature space affects here. We can conclude from the explained variance of 100 % that dimensionality could have been reduced more. We also have to remember that Calinski-Harabasz index might suffer even from moderate noise [26].

We then use Adjusted Rand Index to check the quality of resulting clustering compared to ground truth. ARI tells us the similarity of the defined ground truth clustering and our resulting clustering. Figure 5.2 shows ARI values for different number of clusters $k = [2, 12]$. Because ARI uses the ground truth cluster labels for the publications, it peaks at number of clusters at 3. But we see that its value for $k = [4, 6]$ is almost equally good. We interpret this to mean that each of these clusterings with $k = [3, 6]$ have approximately equal amount of misclustered publications, so our clustering method can't actually find above others clustering for $k = 3$ from this data.

We inspected the resulting clustering for three clusters by looking at the most frequent terms of the publications in each cluster. Top 15 terms for

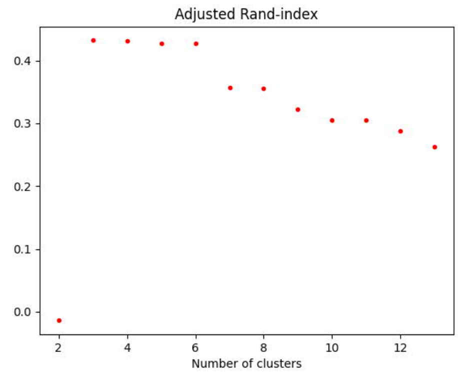

Figure 5.2: Hierarchical clustering. Adjusted Rand index for the manually annotated set of 455 publications clustered with hierarchical clustering, LSA with 800 components. The higher value means more compact and better clustering. The index gains its peak value with number of clusters at 3.

Cluster	**Top terms**
1	service image map system filter user som neural document mobile feature rule self-organizing logic query
2	dementia vascular vad trial subcortical cognitive alzheimer criterion impairment cause lesion stroke subtypes prevalence ad
3	stroke epilepsy pain treatment risk ad child seizure headache cognitive pd cortex parkinson cell apoe

Table 5.1: Top terms for manually annotated dataset of 455 articles

each cluster are shown in the Table 5.1. We can see that there is one cluster with computer science related terms and two clusters with clinical neurology related terms. So we can notice that our clustering method didn't find the clusters we expected. This is also visible in adjusted Rand index values for the number of clusters $k = [3, 6]$ almost equals (Figure 5.2). For three clusters, part of the samples are incorrectly clustered, when compared against manual labelling, and for clusterings with four to six clusters, the situation is obviously the same.

For a comparison we also tried clustering manually annotated data with K-means clustering. The idea was to see if this method would give clearer clusters where the clustering that either silhouette values or Calinski-Harabasz index would recognize. Its similar index values for number of clusters $k = [2, 12]$ can be seen in figure 5.3. K-means wasn't able to significantly improve the clustering.

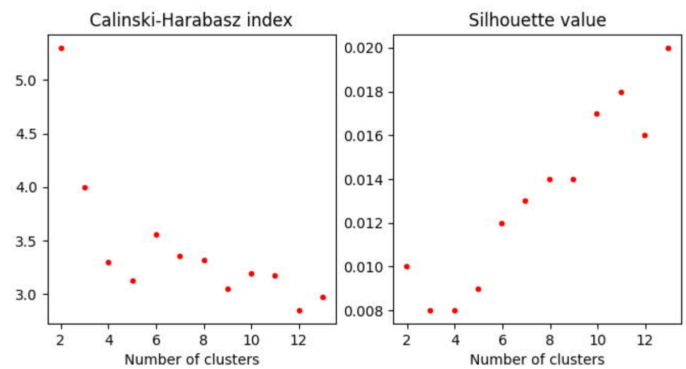

Figure 5.3: Calinski-Harabasz index and average silhouette values for the manually annotated set of 455 publications clustered with *K-means clustering*, LSA with 800 components. The higher value means more compact and better clustering in both graphs.

We can see that both our internal validation methods indicate that our clustering method doesn't find significant clusterings for expected three clusters. As a result we also don't get strong support for using any of these two measures to decide the number of clusters for the actual one year data. We cluster next the full data for years 2000-2001 to see if more data improves the clustering results.

5.2 Finnish publications from years 2000-2001

The analysis of full data set followed the same steps as for the manually annotated data except we couldn't calculate adjusted Rand index because of the lacking ground truth. Vectorising the two years of data, 21155 publications, with term frequency cut-off values $min_d f = 2$ (number of occurrences) and $max_d f = 0.1$ (portion of documents with the term) and maximum number of terms 50000 resulted 101333 ignored terms and a feature space of 50000 most frequent terms just below the $max_d f = 0.1$ threshold. This feature space was then reduced with LSA to 800 component reduced space. In this dimensionality reduction 35 % of the explained variance of the 50000 term feature space data was retained. The data was then clustered using Ward's method with selected values of number of clusters in the range of $k = [4, 500]$ to find the optimal value for k. The Calinski-Harabasz and average silhouette values for this range are shown in Figure 5.4. We can see that the index

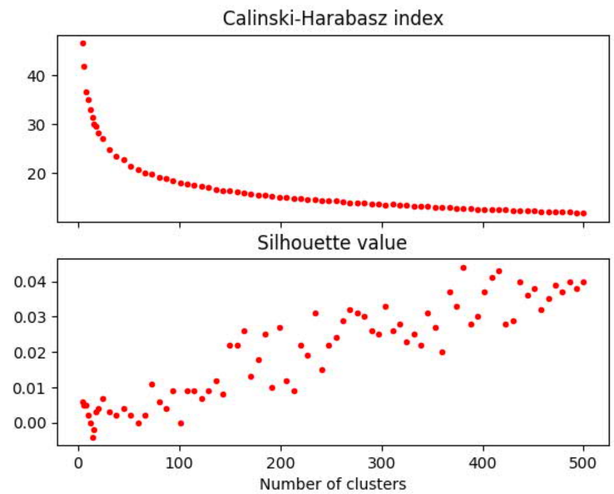

Figure 5.4: Calinski-Harabasz index and average silhouette values for Ward's clusterings the 2000-2001 Finnish publications data with number of clusters $4 - 500$. The higher value mean more compact and better clustering in both graphs.

values follow the same pattern as with the manually annotated data. Again Calinski-Harabasz index suggests that optimal clustering is reached with very few clusters while silhouette values tells us that clusterings improve with increasing number of clusters. Because these cluster validation measures are in conflict we chose a compromise value for the number of clusters, 188, for further inspection. We chose this as a haphazard sample in the order of magnitude of current amount of WoS subject categories (254). We wanted to validate by looking into clusters whether there would be some meaningful clustering.

We first plotted the silhouette of this clustering in Figure 5.5. Colour indicates the cluster. Each cluster consists of stacked and infinitesimally thin horizontal lines. Each line corresponds a silhouette value of a publication in its cluster. From the plot we see that some of the clusters are limited in size and well clustered indicated by thinner horizontal silhouettes extending further to the positive axis direction. There are also few very big clusters of

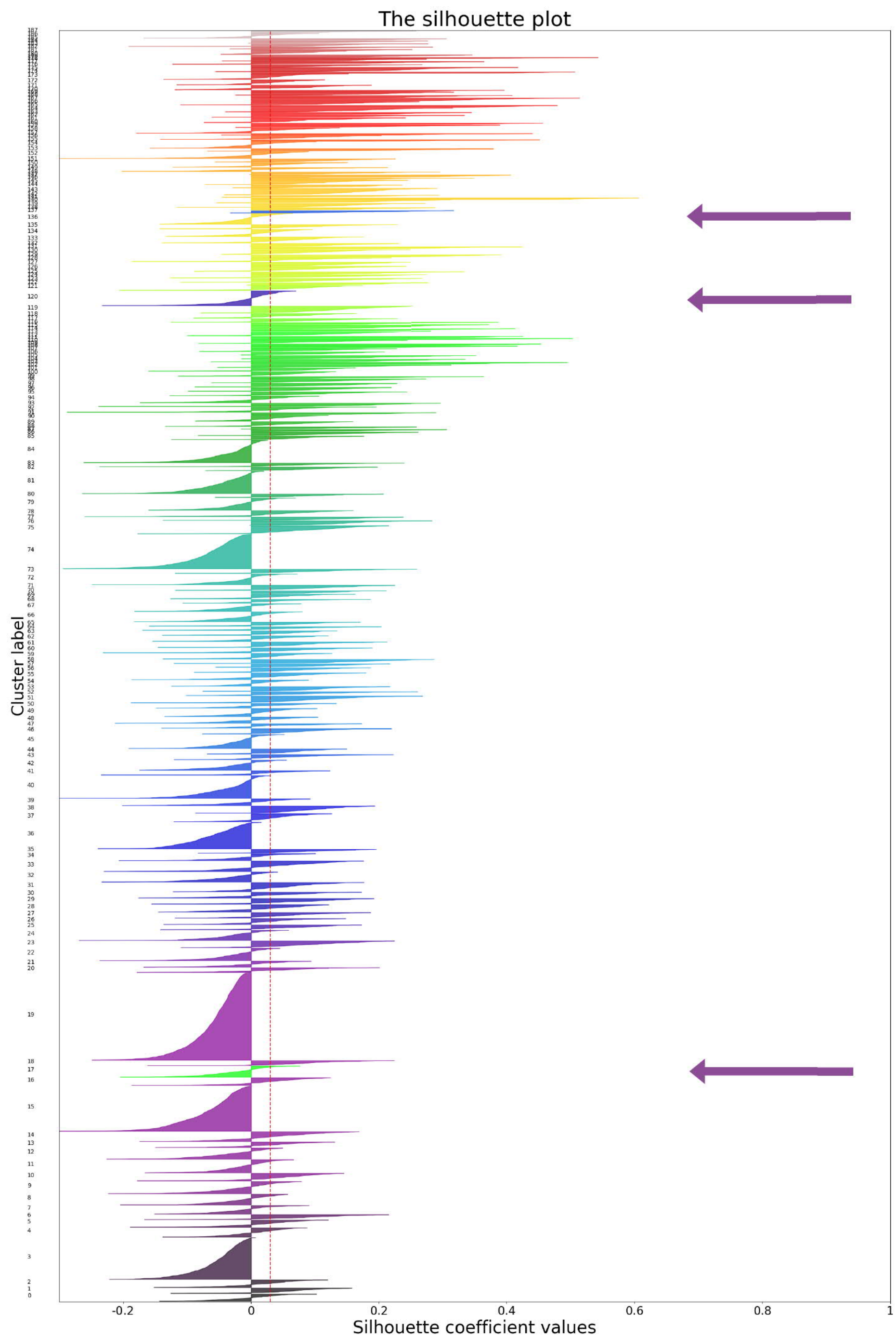

Figure 5.5: Silhouette values of 188 clusters. Repeating shapes are clusters consisting of stacked thin horizontal lines of silhouette values of individual items. Colour indicates the cluster. Contrasting coloured ones indicated with three arrows are the sample clusters. Dashed line is the average over whole clustering.

items poorly clustered among their neighbouring items indicated by broader silhouettes extending to the negative axis direction. Many clusters have both components, resulting a silhouette resembling bow tie. Short lines as part of a silhouette indicate items that could almost equally well be clustered in their nearest neighbouring cluster. As a result the average silhouette value of this clustering 0.030, (dashed line in the plot) is very low in absolute terms considering its index space [-1,1].

By inspecting the top 15 terms of three random clusters sampled from 188 in Table 5.3, we note that at least these three clusters seem reasonable. The

Cluster	Top terms
9	process paper system technology project development product information design action machine quality agent concept environment
19	map som document text self-organizing algorithm content approach application visualization cluster sgml network search structure
31	mg treatment toremifene placebo day month efficacy symptom week receive docetaxel tamoxifen daily breast dose

Table 5.3: Top terms for three random clusters from total 188

cluster 9 contains terms perhaps from some production technology area, the cluster 19 seems to contain computer science related terms and the cluster 31 medical research related terms. To get little more concrete understanding of these clusters, we listed five random publication titles and their WoS subject categories in the Table 5.4. (See Appendices B and C for top term and publication title lists for ten sample clusters respectively.) We can now compare our first impression of these three clusters, given by top terms, to full publication titles and WoS subject categories inherited from the journal in which a publication was published. The cluster 9 turns out to be a mixed cluster of quite diverse publications including political science, psychology, ecology and forestry. The cluster 19 contains computer science related publications but also a publication related to music that could be an outlier. The sample publications of cluster 31 are all related to medical sciences.

Table 5.4: Five random articles for three random clusters out of 188 clusters for the publications of years 2000-2001

Clust	**Title**	**WoS category**
9	Size of environmental grain and resource matching	ECOLOGY
	A note on Matthias Sutter	POLITICAL SCIENCE
	National outdoor recreation demand and supply in Finland: an assessment project	FORESTRY
	Industrial ecology of the paper industry	
	Sex differences in quality of life among allogeneic BMT recipients	PSYCHOLOGY; SOCIAL SCIENCES BIOMEDICAL
19	Neuron weight dynamics in the SOM and Self-Organized Criticality	COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE; ENGINEERING ELECTRICAL ELECTRONIC
	Indexing text with approximate q-grams	COMPUTER SCIENCE THEORY METHODS
	XML based text TV	COMPUTER SCIENCE INFORMATION SYSTEMS; ENGINEERING ELECTRICAL ELECTRONIC; TELECOMMUNICATIONS
	Musical networks: Parallel distributed perception and performance.	PSYCHOLOGY EXPERIMENTAL; MUSIC
	PARNEU: general-purpose partial tree computer	COMPUTER SCIENCE HARDWARE ARCHITECTURE; COMPUTER SCIENCE THEORY METHODS; ENGINEERING ELECTRICAL ELECTRONIC
31	Citalopram controls phobic symptoms in patients with panic disorder: randomized controlled trial	NEUROSCIENCES PSYCHIATRY PSYCHIATRY

	Long-term treatment of psoriasis with calcipotriol scalp solution and cream	DERMATOLOGY VENEREAL DISEASES
	Docetaxel, a promising novel chemotherapeutic agent in advanced breast cancer	ONCOLOGY
	Silica xerogel as an implantable carrier for controlled drug delivery - evaluation of drug distribution and tissue effects after implantation	ENGINEERING BIOMEDICAL; MATERIALS SCIENCE BIOMATERIALS
	A double-blind, randomized study to compare the efficacy and safety of terbinafine (Lamisil (R)) with fluconazole (Diflucan (R)) in the treatment of onychomycosis	DERMATOLOGY VENEREAL DISEASES

In Figure 5.5 these clusters are coloured with contrasting colours, green and blue. (Cluster numbering got mixed, so in the figure they have different positions.) The lowest green cluster corresponds to the cluster 31. Regardeless of the consistent top terms and the sample of publication titles the cluster has its silhouette values in $[-0.2, 0.1]$ range and the negative part of the silhouette is slightly larger than positive. This tells us that the cluster isn't very coherent when half of its items seems to be misclustered. The second blue cluster in the upper part of the plot corresponds to the cluster 9. It has similar silhouette shape with cluster 31 which is in line with the mixed sample of titles it contains. The blue top cluster corresponds to the cluster 19. Its values are in the $[-0.05, 0.35]$ range which promises for a more meaningful cluster. This is in line with the top term list and the sample of titles. We see even from this narrow glimpse that the clustering wasn't satisfactory. While the cluster 31 seemed consistent by top terms and the sample of its titles, the silhouette values tells otherwise. Top terms and sample titles of clusters 9 and 19 are more in line with their silhouette values. Generally these random samples show that clustering like this can't substitute currently used classifications.

Chapter 6

Discussion

The usefulness of our results lies in that they show how difficult the task at hand is, and what aspects there are to consider when approaching the problem. The problem of defining how the science should be classified into different disciplines is quite difficult and ambiguous.

The data was incomplete but we chose to use it as such without pruning the possible defective samples. This was due to the amount of data, manually checking over 20000 titles was not an option and sufficient general rules to filter the data did not occur to us in the beginning of the work. Some minimum length for the publication data could leave out clearly inadequate samples and perhaps improve the clustering results. General outlier detection techniques [20] could prove problematic here when there obviously will be single publications in some disciplines. There probably would be disciplines with very few publications in spite of including more data.

We tried to include a larger data set, covering years 2000-2003, 47303 publications, to cluster. This resulted running out of memory in our laptop environment with 16 GB of RAM. Obviously more capacity would have been available but taking over a new environment was not possible within the limits of this work.

Tokenization during pre-processing left one and two character words such as measurement units in data which was not desired. Adjusting tokenization or extending the stop word list should filter these out. During vectorisation we hit our limit of maximum number of terms of 50000. This might be undesirable because now our terms were from the most frequent end of the frequency range of minimum 2 occurrences – maximum in 10 % of documents. This means that rarer and thus more specific terms might have been discarded. We tested some hand picked values of minimum and maximum document frequency threshold for terms but more systematic exploration of these parameters could give better set of features. Furthermore, the discarded terms contained quite many compounds joined with underscore '_' to conserve the constructs. It might be worth to try and split those compounds to enable the individual words to be counted. On the other hand, using 2-grams or higher order n-grams in addition to single words could also be interesting to try, although it will again increase the number terms.

Selected number of components in reducing dimensionality of the data resulted quite low proportion of explained variance (35 %). On the other hand, the resulting 800 component low-dimensional space is quite high dimensional too. With the manually annotated data set the explained variance was 100 %. Yet the internal validation metrics gave insignificant values. This probably shows that 800 dimensional data with just 455 samples is still too high dimensional [7]. Selecting and extracting features more carefully and then reducing dimensionality more could help. Transforming data somewhere to $[50 - 150]$ components could be more justified range [16]. Overall using more data such as whole world data could improve clustering results.

The model selected, agglomerative Ward's clustering, is one of the basic clustering methods. The method should also tolerate noise somewhat. We selected it because it produces hierarchical cluster structure that is naturally expected of fields and sub fields of science. On the other hand, the existense of cross- disciplinary fields such as bioinformatics suggest that the taxonomy of science could be understood as directed asyclic graph. That is, a tree with each child node having possibly multiple parent nodes and all edges directed such a way that the graph has no cycles. It is known to work well only with compact, even-sized clusters. This might be a problem with the current data where different disciplines are expected to be varying in size. It can't cluster very large data sets because of the computational cost. We might have bumped to that limit in a smaller scale while trying to cluster the four year data of 47000 publications. The selected model is a hard clustering method in a sense that a publication is clustered to one discipline explicitly.

Choosing the number of clusters was a difficult part of this problem. We tried two internal validation methods, Calinski-Harabasz criterion and silhouette values. Neither of them could reveal any meaningful optimal clustering if any occurred. Our manually annotated data set for choosing the number of clusters was probably too small (455) compared to number of components (800). It has been also noted that Calinski-Harabasz index might suffer even from moderate noise [26]. There are other internal clustering validation metrics that might work better with our data, one perhaps worth testing could be S_Dbw (Scatter and Density between clusters) index [18]. It might also be that the merging of two computer science related sub fields into one cluster and clinical neurology splitting into two clusters can also result from the true properties of the data. Perceived human definitions might not always match up with the underlying data. So there might be two sectors of clinical neurology represented in the data that are more distant from each other than these publications related to two computer science fields.

We didn't utilize the hierarchical structure of agglomerative clustering in this work. It could have helped in deciding the number of clusters [24].

We could have very roughly compared our clustering to existing classification by calculating adjusted Rand index against WoS subject categories. The question there would be how to automatically select WoS category for comparison for a publication that has several categories.

Conclusions

We tried a basic agglomerative clustering with Ward's method on Finnish publications from years 2000-2001 to see if meaningful clustering by scientific discipline could be obtained. Some sensible clusters might have been formed but there appeared also totally mixed clusters which reduces the overall result as unsatisfactory. The clustering method was tested to the limit of our laptop environment with the amount of data.

Future work should improve the pre-processing of the data, consider using larger data set and test hybrid models mixing both text and citation analysis. Relaxing the hard clustering assumption to allow for soft clustering of probability based discipline labelling could also prove interesting.

Bibliography

[1] Princeton University "About WordNet." WordNet. Princeton University, 2010.

[2] *Tieteen tila 2018 - Suomen Akatemia, [Status of science 2018 - Academy of Finland]*. 2018.

[3] The 2012 ACM Computing Classification System, 2020.

[4] American Physical Society Physics Subject Headings home page., 2020.

[5] Ministry of Education and Culture, Vipunen - Education Statistics Finland, Bibliometrics (Web of Science) service, 2020.

[6] National Library of Medicine Classification Home Page, 2020.

[7] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory* (Berlin, Heidelberg, Jan. 2001), ICDT '01, Springer-Verlag, pp. 420–434.

[8] Alpaydin, E. *Introduction to Machine Learning*. The MIT Press, Cambridge, 2004.

[9] Archambault, E., Beauchesne, O. H., and Caruso, J. Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th international conference of the international society for scientometrics and informetrics* (2011), 66–77. Durban South Africa.

[10] Bellis, N. D. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Scarecrow Press, Mar. 2009. Google-Books-ID: ma4YjaKyM9cC.

[11] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, Jan (2003), 993–1022.

[12] Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., and Börner, K. Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLOS ONE 6*, 3 (Mar. 2011), e18029.

[13] Broadus, R. Toward a definition of "bibliometrics". *Scientometrics 12*, 5-6 (1987), 373–379.

[14] Calinski, T., and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods 3*, 1 (1974), 1–27.

[15] COLE, F. J., and EALES, N. B. THE HISTORY OF COMPARATIVE ANATOMY: PART I.—A STATISTICAL ANALYSIS OF THE LITERATURE. *Science Progress (1916-1919) 11*, 44 (1917), 578–596.

[16] Dolnicar, S. A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation. *Faculty of Commerce - Papers (Archive)* (Dec. 2002).

[17] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, May 1988), CHI '88, Association for Computing Machinery, pp. 281–285.

[18] Halkidi, M., and Vazirgiannis, M. Clustering validity assessment: finding the optimal partitioning of a data set. In *Proceedings 2001 IEEE International Conference on Data Mining* (Nov. 2001), pp. 187–194.

[19] Hann, M. L. Towards an Algorithmic Methodology of Lemmatization. *Association for Literary and Linguistic Computing Bulletin 3*, 2 (1975), 140–150. Num Pages: 11.

[20] Hodge, V. J., and Austin, J. A survey of outlier detection methodologies. *Artificial Intelligence Review 22* (2004), 2004.

[21] Hubert, L., and Arabie, P. Comparing partitions. *Journal of Classification 2*, 1 (Dec. 1985), 193–218.

[22] Janssens, F., Zhang, L., Moor, B. D., and Glänzel, W. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management 45*, 6 (Nov. 2009), 683–702.

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

[24] Kimes, P. K., Liu, Y., Hayes, D. N., and Marron, J. S. Statistical significance for hierarchical clustering. *Biometrics 73*, 3 (2017), 811–821. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12647.

[25] Korenius, T., Laurikkala, J., Juhola, M., and Järvelin, K. Hierarchical clustering of a Finnish newspaper article collection with graded relevance assessments. *Information Retrieval 9*, 1 (Jan. 2006), 33–53.

[26] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. Understanding of Internal Clustering Validation Measures. In *2010 IEEE International Conference on Data Mining* (Sydney, Australia, Dec. 2010), IEEE, pp. 911–916.

[27] Luhn, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development 1*, 4 (Oct. 1957), 309–317. Conference Name: IBM Journal of Research and Development.

[28] Luhn, H. P. Key word-in-context index for technical literature (kwic index). *American Documentation 11*, 4 (1960), 288–295. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090110403.

[29] Manning, C. D., Raghavan, P., and Schütze, H. Introduction to Information Retrieval, July 2008. ISBN: 9780521865715 9780511809071 Publisher: Cambridge University Press.

[30] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[31] Persson, O., Luukkonen, T., and Hälikkä, S. A Bibliometric Study of Finnish Science. Tech. Rep. 48/00, VTT Technical Research Centre of Finland., Espoo, 2000.

[32] Pritchard, A. Statistical bibliography or bibliometrics? *Journal of Documentation 25* (01 1969), 348–349.

[33] PUDOVKIN, A. I., AND GARFIELD, E. Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology 53*, 13 (Nov. 2002), 1113–1119. Publisher: John Wiley & Sons, Ltd.

[34] ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20* (Nov. 1987), 53–65.

[35] SHU, F., JULIEN, C.-A., ZHANG, L., QIU, J., ZHANG, J., AND LARIVIÈRE, V. Comparing journal and paper level classifications of science. *Journal of Informetrics 13*, 1 (Feb. 2019), 202–225.

[36] SIEMENS, R. Lemmatization and parsing with *TACT* preprocessing programs. *Digital Studies/Le champ numérique*, 1 (Feb. 1996). Number: 1 Publisher: Open Library of Humanities.

[37] STRAUSS, T., AND MALTITZ, M. J. V. Generalising Ward's Method for Use with Manhattan Distances. *PLOS ONE 12*, 1 (Jan. 2017), e0168288. Publisher: Public Library of Science.

[38] SUOMINEN, A., AND TOIVANEN, H. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology 67*, 10 (Oct. 2016), 2464–2476.

[39] TRUNK, G. V. A Problem of Dimensionality: A Simple Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*, 3 (July 1979), 306–307. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[40] WALTMAN, L., AND ECK, N. J. V. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology 63*, 12 (2012), 2378–2392.

[41] WANG, Q., AND WALTMAN, L. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics 10*, 2 (May 2016), 347–364.

[42] WARD, JR, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association 58*, 301 (Mar. 1963), 236–244. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845.

[43] WILLETT, P. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management 24*, 5 (Jan. 1988), 577–597.

Appendix A

Manually annotated data set

Sample of 20 publications of total 455 from manually annotated data set. Discipline is the resulting classification and taken action is in 'Clarification' column.

Title	Discipline	Clarification
Assessing human-centred design processes in product development by using the inuse maturity model	information systems	2nd opinion, manually reclassified
Information quality process - A primary success requirement in management information systems projects: Results from a case study.	information systems	bad data, excluded
Model-based sound synthesis of tanbur, a Turkish long-necked lute	information systems	Classified both 0 and 1, excluded
Birth of a new world order: Ice-Borg	information systems	duplicate, excluded
A characterization of poly-slender context-free languages	information systems	
A cross-cultural study on escalation of commitment behavior in software projects	information systems	
Birth of a new world order: Ice-Borg	artificial intelligence	2nd opinion, excluded
A similarity evaluation technique for data mining with an ensemble of classifiers	artificial intelligence	2nd opinion, manually reclassified

Title	Discipline	Clarification
Arbiter meta-learning with dynamic selection of classifiers and its experimental investigation	artificial intelligence	Classified as 0, manually reclassified
Musical instrument recognition using cepstral coefficients and temporal features	artificial intelligence	Classified both 0 and 1, manually reclassified
Fuzzy interval input-output aggregation of switched traffic in communication networks	artificial intelligence	unclear classification, excluded
A Bayesian weighting principle for the fundamental matrix estimation	artificial intelligence	
A new basis for spreadsheet computing: Interval Solver (TM) for Microsoft (R) excel	artificial intelligence	
Gender differences in associations between religious attendance and mental health in Finland	clinical neurology	2nd opinion, excluded
A father and a disabled child	clinical neurology	bad data, excluded
Supported employment in Russia: Dream or reality?	clinical neurology	bad data, excluded
A longitudinal quantitative EEG study of Alzheimer's disease: Relation to apolipoprotein E polymorphism	clinical neurology	
A population-based study on epilepsy in mentally retarded children	clinical neurology	
Accuracy of pedicle screw insertion with and without computer assistance: a randomised controlled clinical study in 100 consecutive patients	clinical neurology	
Attitudes of Finnish Members of the European Parliament toward inclusion	NA	bad data, excluded

Appendix B

Top terms

Top 15 terms of ten random clusters of total 188 clusters created by Ward's method.

Cluster	Top terms
9	process paper system technology project development product information design action machine quality agent concept environment
19	map som document text self-organizing algorithm content approach application visualization cluster sgml network search structure
31	mg treatment toremifene placebo day month efficacy symptom week receive docetaxel tamoxifen daily breast dose
44	r valuation non-wood benefit easyhaler emphasize forest particularly device turbuhaler dose asthma salbutamol powder inhaler
55	food vitamin intake dietary fat antioxidant supplementation child risk stroke ci diet beta-carotene serum cholesterol
61	temperature heat material degree x diffusion profile stress sample density alloy phase simulation property glass
93	wave resonator ionospheric alfven pc1 ipdp iar frequency artificial ground ionosphere eiscat vertical pearl pulsation
131	to5 kb apart equal gene chromosome candidate finnish linkage locus mutation syndrome association genome disease
143	lignment pigtailing low-cost fiber technique meg magnetocardiographic auditory clinical somatosensory evoke laboratory assessment cognition mcg
144	nicotine dopamine striatal chronic rat da kg withdrawal locomotor dopac nicotinic hva concentration infusion mg

Appendix C

Articles

Titles and existing WoS subject categories of five random publications from ten random clusters of total 188 clusters created by Ward's method.

Clust	Title	WoS category
9	Size of environmental grain and resource matching	ECOLOGY
	A note on Matthias Sutter	POLITICAL SCIENCE
	National outdoor recreation demand and supply in Finland: an assessment project	FORESTRY
	Industrial ecology of the paper industry	
	Sex differences in quality of life among allogeneic BMT recipients	PSYCHOLOGY; SOCIAL SCIENCES BIOMEDICAL
19	Neuron weight dynamics in the SOM and Self-Organized Criticality	COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE; ENGINEERING ELECTRICAL ELECTRONIC
	Indexing text with approximate q-grams	COMPUTER SCIENCE THEORY METHODS
	XML based text TV	COMPUTER SCIENCE INFORMATION SYSTEMS; ENGINEERING ELECTRICAL ELECTRONIC; TELECOMMUNICATIONS

	Musical networks: Parallel distributed perception and performance.	PSYCHOLOGY EXPERIMENTAL; MUSIC
	PARNEU: general-purpose partial tree computer	COMPUTER SCIENCE HARDWARE ARCHITECTURE; COMPUTER SCIENCE THEORY METHODS; ENGINEERING ELECTRICAL ELECTRONIC
31	Citalopram controls phobic symptoms in patients with panic disorder: randomized controlled trial	NEUROSCIENCES PSYCHIATRY PSYCHIATRY
	Long-term treatment of psoriasis with calcipotriol scalp solution and cream	DERMATOLOGY VENEREAL DISEASES
	Docetaxel, a promising novel chemotherapeutic agent in advanced breast cancer	ONCOLOGY
	Silica xerogel as an implantable carrier for controlled drug delivery - evaluation of drug distribution and tissue effects after implantation	ENGINEERING BIOMEDICAL; MATERIALS SCIENCE BIOMATERIALS
	A double-blind, randomized study to compare the efficacy and safety of terbinafine (Lamisil (R)) with fluconazole (Diflucan (R)) in the treatment of onychomycosis	DERMATOLOGY VENEREAL DISEASES
44	Genetics of disease - Away from the beaten track - Editorial overview	BIOCHEMISTRY MOLECULAR BIOLOGY; BIOTECHNOLOGY APPLIED MICROBIOLOGY; CELL BIOLOGY; GENETICS HEREDITY
	Different sensitivities of human and rat rho(1) GABA receptors to extracellular pH	NEUROSCIENCES; PHARMACOLOGY PHARMACY
	Life cycle profit calculations for lubrication	MATERIALS SCIENCE PAPER WOOD

	Wheat and rye brans: What is the difference?	PLANT SCIENCES; FOOD SCIENCE TECHNOLOGY
	Salbutamol via Easyhaler (R) produces equivalent bronchodilation to terbutaline via Turbuhaler (R) following inhalation of a single dose of equipotent beta(2)-sympathomimetic	PHARMACOLOGY PHARMACY
55	Long-term effects of vitamin E, vitamin C, and combined supplementation on urinary 7-hydro-8-oxo-2 '-deoxyguanosine, serum cholesterol oxidation products, and oxidation resistance of lipids in nondepleted men	HEMATOLOGY; PERIPHERAL VASCULAR DISEASE
	Factors affecting broadband ultrasound attenuation results of the calcaneus using a gel-coupled quantitative ultrasound scanning system	ENDOCRINOLOGY METABOLISM
	Controlled trial of alpha-tocopherol and beta-carotene supplements on stroke incidence and mortality in male smokers	HEMATOLOGY; PERIPHERAL VASCULAR DISEASE
	Quercetin intake and the incidence of cerebrovascular disease	NUTRITION DIETETICS
	Regional cerebral blood flow during exposure to food in obese binge eating women	CLINICAL NEUROLOGY; NEUROIMAGING; PSYCHIATRY
61	An engineering model for heating energy and emission assessment - The case of North Karelia, Finland	ENERGY FUELS; ENGINEERING CHEMICAL
	Lattice sites of diffused gold and platinum in epitaxial ZnSe layers	
	Pevotella pallens has a unique phospholipid analogue profile.	DENTISTRY ORAL SURGERY MEDICINE

	Heat transfer enhancement at solid-liquid and solid-gas interfaces by near-surface coolant agitation	ENGINEERING MANUFACTURING; ENGINEERING ELECTRICAL ELECTRONIC; MATERIALS SCIENCE MULTIDISCIPLINARY
	Peaked density profiles: evidence of inward transport in the W7-AS stellarator	PHYSICS FLUIDS PLASMAS; PHYSICS NUCLEAR
93	Non-stationary Alfven resonator: vertical profiles of wave characteristics	GEOCHEMISTRY GEOPHYSICS; METEOROLOGY ATMOSPHERIC SCIENCES
	Unified description of non-diffracting X and Y waves	PHYSICS FLUIDS PLASMAS; PHYSICS MATHEMATICAL
	Generation of artificial magnetic pulsations in the Pc1 frequency range by periodic heating of the Earth's ionosphere: indications of ionospheric Alfven resonator effects	
	2D and 3D models for the cross-talk modeling in acoustic devices: A fast-MoM approach	ACOUSTICS; INSTRUMENTS INSTRUMENTATION; PHYSICS APPLIED
	Non-stationary Alfven resonator: new results on Pc1 pearls and IPDP events	
131	Search for schizophrenia genes in Finnish families reveal a locus on Chromosome 1q	GENETICS HEREDITY
	FMR1 CGG expansion to full mutation: What is the lower limit in premutation females?	GENETICS HEREDITY
	High-throughput gene copy number analysis in 4700 tumors: FISH analysis on tissue microarrays identifies multiple tumor types with amplification of the MB-174 gene, a novel amplified gene originally found in breast cancer.	GENETICS HEREDITY

	DTDST is expressed in developing fetal cartilage but also in a wide variety of other tissues and cell types.	GENETICS HEREDITY
	Relative fitness of women with the mitochondrial DNA mutation 3243AG.	GENETICS HEREDITY
143	MEG and MCG in a clinical environment: BioMag Laboratory, Helsinki	BIOPHYSICS; ENGINEERING BIOMEDICAL; INSTRUMENTS INSTRUMENTATION
	Spatial aspects in magnetocardiographic time domain and spectrotemporal post-infarction arrhythmia risk assessment	BIOPHYSICS; ENGINEERING BIOMEDICAL; INSTRUMENTS INSTRUMENTATION
	High performance magnetically shielded room for clinical measurements	BIOPHYSICS; ENGINEERING BIOMEDICAL; INSTRUMENTS INSTRUMENTATION
	Information theory predicts the magnitude of an MEG response related to the operation of the auditory sensory memory	BIOPHYSICS; ENGINEERING BIOMEDICAL; INSTRUMENTS INSTRUMENTATION
	Global optimization in the localization of brain activity	BIOPHYSICS; ENGINEERING BIOMEDICAL; INSTRUMENTS INSTRUMENTATION
144	Cotinine and nicotine inhibit each other's calcium responses in bovine chromaffin cells	PHARMACOLOGY PHARMACY; TOXICOLOGY
	Comparison of the effects of epibatidine and nicotine on the output of dopamine in the dorsal and ventral striatum of freely-moving rats	PHARMACOLOGY PHARMACY
	Effect of acute nicotine administration on striatal dopamine output and metabolism in rats kept at different ambient temperatures	BIOCHEMISTRY MOLECULAR BIOLOGY; PHARMACOLOGY PHARMACY

	The effects of acute nicotine on the body temperature and striatal dopamine metabolism of mice during chronic nicotine infusion	NEUROSCIENCES
	Nicotine-evoked exocytosis from bovine chromaffin cells is independent of phospholipase D activation	NEUROSCIENCES