CLASSIFICATION OF WEB RESIDENT SENSOR RESOURCES USING LATENT SEMANTIC INDEXING AND ONTOLOGIES

Wabo Majavu

A Dissertation submitted to the Faculty of Engineering and the Built Environment,
University of the Witwatersrand, in fulfilment of the requirements of the degree of
Master of Science in Engineering

Johannesburg 2009

CLASSIFICATION OF WEB RESIDENT SENSOR RESOURCES USING LATENT SEMANTIC INDEXING AND ONTOLOGIES

Approved by: ______________________________

Date Approved: ___________________________

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in Engineering in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

Wabo Majavu

This day of 2009

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof.Tshilidzi Marwala for his guidance and academic support as well as his support towards my career and interests. Thank you for your supervision and constant encouragenment to perform better. I would like to thank my mentor Terence van Zyl for his immeasurable contributions. His ideas and disscussions in his office have bloomed to be a fruitful work and I have learnt to say: I was sometimes wrong and Terence was mostly right. Without him this dissertation structure would have never materialised. I wish to thank my mom and my family for their encouragement. I would like to thank God for through Him all is possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Web resident sensor resource discovery plays a crucial role in the realisation of the Sensor Web. The vision of the Sensor Web is to create a web of sensors that can be manipulated and discovered in real time. A current research challenge in the sensor web is the discovery of relevant web sensor resources. The proposed approach towards solving the discovery problem is to implement a modified Latent Semantic Indexing(LSI) by making use of an ontology for classifying Web Resident Resources found in geospatial web portals. This research introduces a new method aimed at improving an information retrieval algorithm, influencing the vector decomposition by including a formal representation of the knowledge of the domain of interest. The aim is to bias the retrieval to better classify the resources of interest. The proposed method uses the domain knowledge, expressed in the ontology to improve the knowledge extraction by using the concept definitions and relationships in the ontology to create semantic links between documents. The clusters formed using the modified algorithm are analysed and performance measured by evaluating the inter-cluster distances and similarity measures within each cluster. The distances are expressed as Euclidean distances of vectors in n-dimensional latent space. The research focus is on investigating how the prior domain knowledge improves the clustering when k-means is used as the partitioning algorithm. It is observed that the modified extraction algorithm can isolate a group of documents that are used to populate the knowledge base, therefore resulting in improved storage of the documents that occur in the geospatial portal. Results found using the combination of ontology and LSI show that clusters are better separated and homogeneous clusters of more specific themes can be formed by hierarchical clustering.

CHAPTER I

INTRODUCTION

1.1 Background Study

Users that require access to geospatial data available via the World Wide Web (WWW), for the purpose of earth observation and environmental modelling, may also require access to information derived from sensor data. The efficiency with which a user can access data can be expressed as a measure of two factors: the search time for the relevant data to be found and returned, and the relevancy ranking of the results where the more relevant the resource, the higher the ranking[1]. The growing volumes of sensor data sets published on the Internet are difficult to locate and access because of the unstructured nature[4]. Examples of these data sets include disaster management and environmental quality assessments. There have been efforts towards the structuring of data and resource description in order to aid discovery, by way of digital libraries, catalogues and resource description and definition[20].

The web is full of sensor resources which when adequately described and catalogued can reduce both search time and accuracy of the returned results. The worldwide web consortium (W3C) refers to a resource as any electronic entity which presents itself as a document, image, service, database or downloadable file[5], while The Internet Engineering Task Force(IETF) refers to a resource as anything that has an identity[7]. The Open Geospatial Consortium (OGC) defines a sensor as a device that responds to a stimulus, such as heat, light or pressure and generates a signal that can be measured or interpreted[42]. Web resident sensor resources (WRSR) are therefore understood to be web accessible electronic entities where sensor data or information about the sensor can be accessed. WRSR can present themselves as sensor data or as a means of accessing sensor data, such as web portals or portal of portals, for example the Global Change Master Directory(GCMD) and the Group on Earth Observation(GEO) portal. Both these portals are dedicated to developing access to a collection of geospatial data taken from different measuring platforms worldwide.

The high irrelevant search results illustrate that current indexing and keyword retrieval mechanisms are insufficient for organising high volumes of data [43]. A sensor web is a global sensor system, connecting sensors and sensor observations world wide in an intelligent and collaborative manner. It is a platform for interconnecting monitoring systems that consist of remote, insitu, dynamic and stationary sensors[35]. One of the goals of the sensor web is to manage and abstract meaningful information from large numbers of sensor systems located worldwide. The OGC Sensor Web Enablement(SWE) initiative is described as a web solution for integrating multiple sensor networks, which currently operate under isolated environments, in order to create a distributed computing infrastructure where sensors and sensor networks can be accessible over the Internet[9]. Discovery of WRSR play a major role in the realisation of the Sensor Web where the vision is to create an automated web of sensors that can be manipulated and discovered in real time[17]. A step towards realising the Sensor Web is the ability to dynamically discover sensor resources published on the WWW and add them to the sensor web.

1.2 Problem Statement

The discovery of sensor resources is a growing problem as an increase in the number of published sensor resources will result in information overload [2]. The problem that is faced with current resource discovery tools is the lack of standardised resource description mechanism, resulting in a limitation of the capabilities of current developed data set and data services discovery applications. The callenge the user faces is high volumes of irrelevant document returned by keyword searches. A more organised structure and document groupings is required. Intelligent systems attempt to address the challenge of relevant data discovery in large data sets made available in such a dynamic environment as the internet. Automatic data classifiers have been used to generate content description of documents found on the web, such as the automatic Resource Description Framework(RDF) for resource discovery, to enable relevant content retrieval from an indexed library[25].

Even with adequate document content description, there is a need to retrieve the documents relevant to the user. Information retrieval(IR) techniques have been explored in data mining research to enhance user searches and structuring of documents based on content[22]. Intelligent systems have been developed for information retrieval to address the inadequacies of performing lexical searches, which involve string matching of the term, which analyse the pattern of usage of a term and the context[8]. Research issues around information extraction and categorisation of structured and semi structured data, have also been widely addressed in literature. The degree of structure is determined by the descriptions of data type accompanying the data[6][24].Therefore when looking at solving the problem of sensor resource discovery the focus is in both sensor descriptions and the challenges of relevant resource content retrieval.

1.2.1 Sensor Data Description

A sensor resource description is a detailed description of the sensor which takes the measurements and a description of the feature which is being observed. Geospatial interoperability standards have been developed to aid in the description and access to sensor data. The sensor web enablement initiative(SWE) has developed a standard for encoding of observations from a sensor known as the Observations and Measurements in order to aid discovery. The problem is the discovery of existing data which has not been encoded in this standard format[13]. Another initiative towards data encodings is the SensorML (Sensor Modelling Language) for describing sensor systems and sensor processes. The standard allows for the description of sensor platform, sensor location and location of observations and sensor properties that can be accessed during tasking[9]. The goal of SWE is to create an open platform for accessing web connected sensors that enable to information such as: flood monitoring, satellite imagery for weather prediction and air quality monitoring[59]. The existing standards however, are the means by which the publishers of the information can make their data discoverable with a focus on descriptions of control interfaces for developing applications that will allow real time data access via the Internet[17]. There is still a need to search for and categorise existing sensor resources that are published. Geospatial portals are the starting point for the task of sensor metadata automatic retrieval and categorisation. The methods adopted for this task can later be implemented for sensor resource web searches and categorising.

1.2.2 Discovery Challenges of Sensor Data

The need to discover sensor resources arises as a result of the need for accessing sensed data that may be used in decision support. There is a large number of sensor resources published by various members of the geoscience community that have been published through varying interfaces. The challenges of information extraction of sensor data are as follows:

- **Data mining challenges**

Current unsupervised data mining techniques such as statistics, neighbourhood and clustering, provide a generic solution to the problem. There is a need for a technique which is tailored to geospatial data and its heterogeneous nature. There arises a need to capture the sensor knowledge and the observations which are expected. This would result in a guided data mining method that is not completely blind, as is the case with general domain non specific unsupervised techniques.

- **Single sensor inventory**

Different organisations publish sensor data across the globe in a manner which best suits their purpose. Other organisations may also deploy sensors with no knowledge that similar sensor systems with the same purpose may exist for a common area. Addressing the challenges of sharing of data will enable better understanding and modelling of the environment.

- **Information about the sensor network**

Varying communication protocols can be used in one network. A description of the sensor network and the environment in which the observations were obtained proves useful when the user requires provenance of the data generated. Completed descriptions are required in an open distributed environment. There is a challenge in publishing large data sets of different types from different sensor networks and include descriptions of each for semantic searches.

- **Imagery and downloadable files**

Intelligent crawling mechanism depend on file extension to be able to classify a resource as containing image content. In the case of downloadable files, user intervention is still required.

1.3 Objectives

The aim is to make web resident sensor resources more discoverable by classification of unstructured web pages that can be accessed through the portal according to measured features using text classification techniques. Text classification involves categorising unknown documents, the portal web pages, into categories or classes and can either be supervised or unsupervised. An example of unsupervised text classification is clustering, which is the grouping of objects with similar attributes into clusters. These techniques have previously been used in natural language classification and are to be adopted for classification of the web pages in combination with ontologies. A portion of the web, web sensor portals, is chosen as the starting point for sensor resource classification and discovery.

1.4 Research Approach

The hypotheses is that a modification to the LSI algorithm can adapt it to better model sensor resources published in a geospatial portal. This was tested by evaluating the relevance with regards to theme modelling of each web page found in the portal. First the algorithm was tested on the Global Change Master Directory (GCMD) portal where data categories are already known and then later implemented for sensor resource modelling with respect to extracted feature, by using the classified documents to extract sensor information. The aim of constructing an ontology was to capture the domain knowledge by defining concepts and expressing their relationships. The combination of the ontology with the Information Retrieval algorithm, LSI, was evaluated to compare the document groupings according to the relevant themes that were extracted. The strength of the technique used lies in the introduction of prior knowledge to the information extraction process, with the aim of producing improved clusters that better represent the document groupings and the information found on each document. The two methods that are compared are the original LSI algorithm and the modified LSI-ontology. The test environment is the documents found in the portal. This is a starting point that can later be applied to web searches of sensor data.

1.5 Overview of Results

The results of the research has revealed that there is an improvement in the clustering by using LSI-ontology. The performance is determined by evaluating the distances of the clusters from each other and within each cluster. It is found that the clusters can further be subclustered in order to narrow down the scope of each theme, in order to create instances for ontology population. The purpose of the knowledge base population is to capture the portal structure and information regarding the document groupings to better enable searches in the future. Once the documents are clustered then the user can access the document in order to download the sensor data.

1.6 Thesis Outline

In chapter 2, the background information is presented relating to resource discovery within the sensor web. Chapter 3 presents an overview of text mining techniques and how ontologies have been used previously in combination with information retrieval. A description of LSI is also presented, followed by a description of the Web Resident Sensor Resource Ontology. Chapter 4 is the method followed in gathering data and carrying out of the experiment. The results are presented in Chapter 5 where analysis and discussion is covered at length. Finally Chapter 6 presents the conclusion, critical evaluation of the work done and recommendations for future research.

CHAPTER II

SENSOR RESOURCE DISCOVERY IN THE SENSOR WEB

The problem statement in Chapter 1 presents sensor resource discovery as a problem to be addressed and the sensor web as the application environment. In this chapter the related work in these areas is discussed. In the area of resource discovery, a description of current discovery mechanisms used is presented for unorganised and unstructured environments, such as the World Wide Web, also referred to as first time discovery. Discovery is then described for structured environments, where the resources have been collected and arranged to enable retrieval. With respect to the Sensor Web, a definition and current initiatives, followed by the challenges is given. Finally, some of the approaches to resource discovery within the Sensor Web are discussed, with a focus on the use of ontologies and a description of sensor document clustering.

2.1 Sensor Resource Discovery

Sensor resource discovery refers to the discovery of sensor resources that are available on the internet. In general, resource discovery can be classified into two different scenarios: one where new sensor resources are being discovered for the first time, the other where discovery takes place in a catalogued storage medium, eg. indexes, centralised registries or libraries[58]. The mechanisms to perform these two types of resource discovery differ, where first time discovery is performed in an unstructured environment and the latter in one that is structured. Unstructured discovery is usually performed in order to create and provide descriptions for the resources discovered during a search and a structured set contains descriptions or annotations and allows for queries to be carried out, usually with a higher search precision, which is determined by the relevance factor of the results to a search[29].

2.2 Discovery in unstructured environment

An overview of the most common techniques that are widely used for first time discovery of resources is as follows:

- **Web Crawling**

Significant work has been done in the field of web crawling as a method to discover published resources. Crawlers or spiders are programs that collect web pages and follow links on each page in order for indexing to be done by search engines. Different techniques have been implemented in particular for topic specific crawling in an attempt, amongst others, to create topic related indexes of search engines to increase both web coverage and query matching of relevant keywords[10]. A focused crawler is defined as a resource discovery systems that seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the web[11]. The documents collected by a web crawler are indexed and stored in a hash table, where keywords can be used to retrieve documents of interest. A search engine is the user interface to access these documents. The results are returned in order, according to a ranking algorithm. Ranking is determined by the number of documents referencing another document, the more times a single document is referred to by others, the higher the ranking of its importance.

- **Browsing**

Browsing involves the user sifting through given information space in search of the relevant content[50] or simply scanning an information space[55]. This method of locating resources is dependant on how organised the resource space is and the amount of information available, therefore cannot be applied to the web at large. A popular browsing method is to traverse a search space by hyperlink navigation, where a user clicks on the links of interest either in a depth first or breath first approach[52]. A more automated mechanism is required for navigating large information spaces, where relational association can be introduced to the tree of topics found on the web or web portal.

- **Searching**

Searching is an automated process, where the user provides some information about the resources being sought,and the system locates some appropriate matches usually by using keyword matches. The scope of the search is dependant on the search engine being used and the parameters given when performing the enquiry. This is a semi structured mechanism as there would already exist and indexed space through which the resource is to be located, in addition to it being constantly populated by the first time discoveries. This technique is often helpful, when compared to browsing, because the information space is reduced and the search performed by the engine is rule based[50]. The different types of search algorithms include breadth first, depth first, random search and uniform cost search[40]. Crawlers alone and search engines do not solve the automatic resource discovery problem, because of the rate of growth of the internet, compared to the rate at which a published web page is published[19]. The rate at which the documents are published far exceeds the rate at which the search engine data base or collection of indexed documents is updated.

2.3 Discovery in structured environment

A structured environment can be described as either a centralised or distributed framework that consists of a database and a query processing engine[52]. Various organisational structures of documents exist, of interest to this work being grouping of documents with similar attributes that characterise subject matter[28]. Classification algorithms have previously been combined with web crawling for the creation of semantic indexes to facilitate speedy and relevant retrieval. Navigation techniques that are deployed in search of web resources for the purpose of creating organised storage, are summarised into: subject gateways, where experts collect domain specific information and arrange it in a catalogue, and web crawlers and search engines which are robots that scour the web in search of indexable documents[49]. Web portals and cataloguing allows for resource metadata to be stored in a database and discovery is performed by searching through the catalogue where a hash table of keys and values is used for matching the user query. These catalogues are static approaches to the problem of resource discovery and often require user intervention, for this reason an automated approach is considered.

2.3.1 Thematic document classification

Data classification is applied in the domain of data mining where the process of searching for a pattern in data is automated and certain rules applied to categorise the unknown data of concern into a class or group membership based on a set of attributes that are tested. Supervised classification requires for the classes belonging to the training data to be known and fully labelled or features extracted. Supervised classifiers can be further broken down into parametric and non parametric classifiers, where parametric classifiers assume a probability distribution for each class and the parameters of the distribution function are estimated, a common example being the Bayes family of classifiers. Non parametric classifiers do not assume a distribution, an error probability is estimated from the training data. With unsupervised classification however, the classes are not predetermined and the clustering or grouping of data with similar attributes is performed blindly. The generic unsupervised classification model uses pattern representation, extracts features or structure from the input data and forms clusters. There are different ways of organising information resources on the web to aid discovery, with clustering being the most common way of performing unsupervised learning.

- **Web Portals**

An example of thematically structured environment are domain specific web portals. Web portals are web sites that collect information on a common topic. They provide links to other web resources and serve as a starting point or a common gateway to networked information services residing on the web[34]. The advantage of domain specific web portals is the single point of access they provide to web services, applications and data published by separate data providers[21]. Even with integrated data access solutions such as portals, data clearing houses and catalogues, there is still a need for structured information representation to enable semantic searches of web resources[39] because portal catalogs are still limited to keyword based retrieval or theme navigation, where users are required to understand how the nested components and links are arranged for each portal.

- **Resource Description**

Resource description plays a key role in resource discovery, enabling easier location of the resource[33]. A resource is described by identifying web resources with Uniform Resource Identifiers (URIs) and indicating relationships among them. Resource description Framework (RDF) is a standardised machine understandable structure, proposed by the the World Wide Web Consortium (W3C) in order to incorporate information about a resource, such as: what type of resource it is, what the subject matter is without accessing and analysing each web page individually, thus enabling discovery[25].

2.4 Resource Classification for sensor discovery

Resource classification has been extensively researched in the field of Information Retrieval (IR) for search optimisation and logical organisation of web documents[30]. Techniques have been explored for sensor resource classification which look at web page classification based on relevant content retrieval[3].Classification of earth and space data by domain conceptualisation has been proposed to aid knowledge discovery in the large amounts of geospatial data available on the web[59].

2.4.1 Sensor Web

The Sensor Web is defined as a web of interconnected sensors which are fronted by interoperable Web Services that comply with standard specifications[17]. The Open Geospatial Consortium (OGC) has proposed a framework of open standards to help enable access and publishing of sensor data across different platforms worldwide as one of the efforts to enabling the creation of the Sensor Web. The realisation of the Sensor Web will facilitate internet accessibility and taskability of heterogeneous web resident sensor resources (WRSR). The Sensor Web operates as an autonomous macro instrument for environmental data gathering and processing which allows for reuse of this collected data for different purposes[16]. One of the functions of the Sensor Web is to allow the discovery of sensor systems and sensor observations filtered according to requirements that are entered by a user[9][41]. Activities that contribute towards the development of the Sensor Web aim to integrate and manage WRSR data. Discovery is targeted at these web resources.

2.4.2 Discovery within the Sensor Web

One of the focuses of the Sensor Web Enablement initiative is the discovery of sensor assets by proposing standard encodings and XML schemas for publishing sensor observations and the description of processes and sensor capabilities[9]. Semantic web technologies have also been used for describing sensor data with spatial, temporal, and thematic metadata.

- **Ontologies**

An ontology is a formal description of concepts and a relation between them, that represents an area of knowledge and is usually expressed in a logic based language. Ontologies together with class instances form a knowledge base which models a domain [44]. Ontologies can be used to share a common understanding of a domain structure and can be used to improve Web based applications by reuse of domain knowledge [5]. Ontologies are used for conceptual modelling in order to define concepts and relationships between concepts with varying levels of expressiveness by allowing restrictions to be placed on property fillers [38]. Semantic technologies make use of Ontologies for knowledge sharing and reasoning [23] and have received attention in the Semantic Web Community as a solution to overcome keyword based information retrieval, where domain specific knowledge bases are created to facilitate semantic searches[38].

The use of Ontologies within the Sensor Web has been explored as a machine readable representation of shared concepts that allows for a mediation between different applications and systems of the sensor web. Sirin used an Ontology to build a Sensor Web Service description[53]. Ontologies in Sensor Web applications is also addressed in the SWAP architecture that proposes a framework for Sensor Web applications being deployed on the internet. This architecture also looks at the use of Ontologies within the Sensor Web to aid semantic sensor resource description and discovery[41].Ontologies can be used as a representation to contextualise sensor resource description within the Sensor Web to aid semantic discovery and retrieval of sensor resources.

2.4.3 Document clustering

Document clustering is grouping of similar documents with similar characteristics. Different clustering algorithms determine how the clusters are formed. Sensor documents can be clustered according to the feature of observation, also known as the theme. The similarity between documents can be computed based on the pattern of usage of words describing the observation and measurements. The characteristics of a cluster are:

- **Centroid**

Every cluster has a centroid, also known as the cluster centre. Clusters are formed so as to minimise the distance between all the cluster members and the centre of the cluster.

- **Threshold**

A document cluster value of similarity is computed, based on the distance from all the cluster centroids. A threshold or minimum distance value is used to determine cluster membership, where if the threshold is exceeded the object will be shifted to another cluster, or a new cluster is formed.

- **Cluster Seed**

A document or object is randomly picked as a cluster member, forming the seed or cluster initiator. The cluster seed is compared to every other incoming document. The seed member is replaced, through an iterative process, by the document with the least average distance to other cluster members.

2.5 Conclusion

This chapter presented background material related to resource discovery within the Sensor Web. Resource discovery is described for structured and unstructured environments and an overview of the mechanisms used in both scenarios is given. Structured environments are described and of particular interest: web portals, where web pages are grouped thematically, or according to searchable spatial coordinates. Sensor Web Portals are used in this research as the test discovery environment and more detail with regard to retrieval mechanisms and clustering methods applied in the portals are given in Chapter 3.

CHAPTER III

INFORMATION RETRIEVAL FOR RESOURCE CLASSIFICATION

Extracting meaning from the web pages found in a geospatial portal requires the understanding of the web page content, which requires automated text processing techniques to be applied to the input data. Text mining makes use of Information Retrieval techniques for the extraction of a pattern from a given set of text containing documents and can be broken down into automatic text classification, topic extraction and trend analysis of topics[27]. With the growing volumes of the Web, significant work is done in the field of text mining and applying natural language processing (NLP) to extract more accurate meaning and representation of text, generally referred to as Information Extraction. In this Chapter we present an overview of text mining and focus on Latent Semantic Indexing (LSI) which is significant to this work. First a discussion is presented on text mining, highlighting some Information Retrieval techniques commonly used for Web documents, followed by a discussion of LSI.

3.1 Text Mining Overview

Text mining is the deriving of meaningful information from text and uses techniques from Information Retrieval, statistics and Machine Learning[27]. The information extracted using these techniques is used to analyse and assess document content. Text mining can be divided into:

- Automatic text classification and clustering

  Text categorisation is the sorting of documents into groups or categories by using either supervised or unsupervised learning methods[57].The documents can be arranged into pre-determined categories or arranged into clusters of similar content. The process combines both information retrieval, extracting and representing meaning and machine learning techniques for building the classifier or clustering mechanism[51].

- Topic extraction

  Detecting topics related to a specified topic or area of knowledge, also referred to as topic discovery and topic spotting. This process involves recognising key topics and fusing or associating similar topics[36].

- Topic trend analysis

  Analysing the number of documents per topic to study the importance or dominance of a particular topic and the pattern of occurrence of the topic for particular text streams[54][47].

From the application areas mentioned above, of particular interest in this study is Automatic text classification and clustering. Text classification has been studied extensively in the field of text and data mining and support vector machines (SVM) and Neural Networks have been used extensively to solve the linear and non-linear classification problem[27]. However, our interest was in using an algorithm to perform unsupervised clustering where there are no pre-determined classes and one that incorporates dimension reduction of the thousands of terms that could be contained in a corpus. The second criteria was to make use of an algorithm that is 'context aware' to incorporate synonyms (different words with the same meaning) and polysemy (same word with different meaning, depending on context). Document content representation is further broken down into feature selection and feature extraction. Feature extraction is used more commonly than feature selection techniques to solve the problem of reducing the dimension, both for computational complexity and also for addressing the synonym problem. The most widely used algorithm with all these properties is Latent Semantic Indexing (LSI) which derives a smaller semantic space to represent the original document. The aim is to apply LSI to the problem of resource discovery on the Web, therefore we look at text classification for the purpose of information retrieval on the Web.

3.1.1 Information Retrieval on the Web

Information Retrieval techniques are used in text mining for organising and locating documents. Concept-based information retrieval is of particular interest because it addresses the problem of searching for information objects and associating meaning to them, rather than with keyword searches. Information retrieval (IR) mechanisms are used to address the problem of relevant data discovery in large document sets. In the case of domain specific retrieval, a need arises to combine conventional IR with domain knowledge in order to make use of the semantics of the data for a guided content extraction. An ontology-driven approach for associating documents and meaning to terms is studied.

3.1.2 Combining Information Retrieval with Ontologies

Ontologies, which have previously been defined as a formal description of concepts and a relation between them to represent an area of knowledge are combined with Information Retrieval techniques for search optimisation. This approach allows for additional relations to be used when a query is entered, for example spatial and temporal relations. The benefit of this approach is to recognise different versions of concept which could belong to the same class having similar properties.[26]. An example of IR combined with the use of ontologies is the ontology based personalised search, which profiles a user query in the form of a concept hierarchy or ontology in order to represent the meaning of the query an then performing an enquiry over a knowledge base.The information objects are stored with their content descriptions and retrieval is performed by matching the user query against the stored objects. This method focuses on information retrieval in an existing static knowledge base, with the assumption that web documents are semantically annotated[46]. A project which is currently underway in developing a semantic framework for web content management using ontologies is the ontology based Web retrieval (ONTOWEB). This system allows semantic searches to be performed in subject gateways and portals to enhance portal navigation. This system, however, requires manual loading of the resource description onto the database and the ontology is used to semantically structure this information which is later queried. In our approach we combine ontology with LSI, a description of which is presented in the following section.

3.2 Description of Latent Semantic Indexing

Latent Semantic Indexing (LSI) was proposed by Deerwester as a document search and retrieval algorithm that organises documents into a semantic structure by using higher order term clustering of the terms contained in the document corpus [14]. It is commonly used as an automatic document classification method and used to measure relevance during searches. The documents are indexed terms and are represented in a semantic space and cosine similarity can be established between documents and a query, without the appearance of a common term. As an automatic indexing and document retrieval method, LSI uses term weighting and normalisation techniques and therefore can be viewed as an extension of the common approaches term frequency and inverse document (TFIDF) approaches.

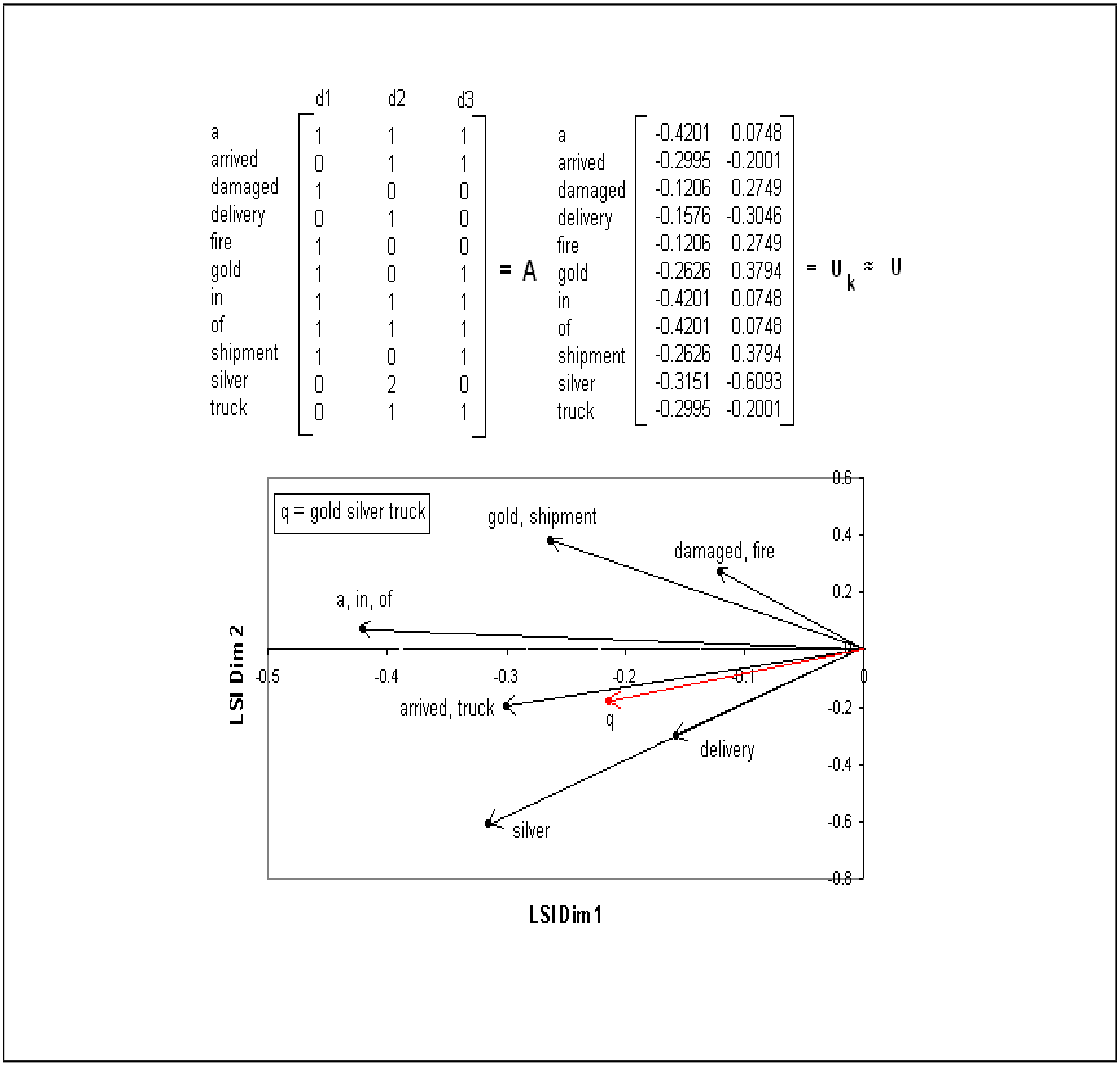
TFIDF weighting computes a vector component for each term, relative to each document. The term document components in the matrix are computed using:

$$W_{t,d} = tf_{t,d} \times log\frac{N}{df_t} \tag{3.1}$$

where, $W_{t,d}$ is the term document weighting, $tf_{t,d}$ the frequency of term t on document d and df_t the number of documents that contain the term t, if N is the total number of documents in the corpus. From the equation it can be seen that the more a term appears on different documents the less weighting will be attributed to the Inverse Document Frequency(IDF) factor so an important word may be discarded because of its frequency throughout the corpus or its infrequency in one document. This method automatically regards a term t as noisy as IDF(t) approaches zero. The unique feature of LSI is to be able to differentiate between relevant words and noisy terms by vector space reduction into relevant categories and the importance of each word computed with respect to each category. The structuring of documents in this manner allows for synonyms to be grouped together because of the underlying pattern of usage of terms. LSI has been implemented successfully[21] in various domains as a semantic annotation, indexing, document retrieval tool.The strength of LSI lies in redistribution of weights across connectivity paths established between documents[39]. Figure 3.1 shows a 2-dimensional implementation of LSI and shows how the terms are indexed according to frequency of appearance on documents d1,d2 and d3. The resulting term-document matrix $\mathbf{A}$ is decomposed into $\mathbf{U}_k$. The plot of term-document clustering is also shown, resulting in 2-dimension coordinates of terms and documents.

3.2.1 Vector Space Representation

After term-weighting, LSI uses a term-document matrix to identify the occurrence of terms and establish a document-term relation within a set of documents. This part of the process incorporates trend analysis and from the matrix it can be determined which terms are more important than others in a corpus. Vector space representation is the matrix where the rows represent terms and the columns represent document indexes. Equation 3.2 shows a matrix $\mathbf{X}$ with dimensions $(m \times n)$, where rows



$$\begin{array}{l|ccc|}
 & d1 & d2 & d3 \\
a & 1 & 1 & 1 \\
arrived & 0 & 1 & 1 \\
damaged & 1 & 0 & 0 \\
delivery & 0 & 1 & 0 \\
fire & 1 & 0 & 0 \\
gold & 1 & 0 & 1 \\
in & 1 & 1 & 1 \\
of & 1 & 1 & 1 \\
shipment & 1 & 0 & 1 \\
silver & 0 & 2 & 0 \\
truck & 0 & 1 & 1
\end{array} = A \quad
\begin{array}{l|cc|}
a & -0.4201 & 0.0748 \\
arrived & -0.2995 & -0.2001 \\
damaged & -0.1206 & 0.2749 \\
delivery & -0.1576 & -0.3046 \\
fire & -0.1206 & 0.2749 \\
gold & -0.2626 & 0.3794 \\
in & -0.4201 & 0.0748 \\
of & -0.4201 & 0.0748 \\
shipment & -0.2626 & 0.3794 \\
silver & -0.3151 & -0.6093 \\
truck & -0.2995 & -0.2001
\end{array} = U_k \approx U$$

Figure 3.1: Term-document 2-dimensional visualisation

represent documents and columns represent terms[48].

$$\mathbf{X} = \begin{bmatrix} d_1t_1 & d_1t_2 \cdots d_1t_n \\ d_2t_1 & d_2t_2 \cdots d_2t_n \\ \vdots & \vdots \\ d_mt_1 & d_mt_2 \cdots d_mt_n \end{bmatrix} \tag{3.2}$$

The matrix is a k-dimension space, and the values represent the weight of term n in document m. The arrangement of terms into vector space ignores grammatical word arrangement by forming a bag-of-words weighted by frequency of occurrence. Vector modelling is used, mostly in keyword searches where a user query is treated as a document with entered text as the terms and a match is made against the bag of words to return relevant documents.

3.2.2 Singular Value Decomposition

Singular Value Decomposition (SVD) is used in the field of linear algebra for the factorisation of a rectangular real or complex matrix in order to establish linear combinations of two sets and express a mathematical relation or correlation that would otherwise not be evident in the original matrix[12]. The first step is to index frequently occurring terms in a term-document matrix and compute singular value decomposition (SVD) from the original k-dimensional term-document matrix. SVD is a matrix decomposition method commonly used for data analysis. The original term-document matrix, **X**, is decomposed into several matrices so their features can be revealed, for example document-document relationships. The decomposition is expressed as,

$$\mathbf{X}(SVD) = \mathbf{T}_{t\times k} \cdot \mathbf{S}_{k\times k} \cdot \mathbf{D}_{k\times d} \tag{3.3}$$

where, T is a left singular vector representing a term by dimension matrix, S is a singular value dimension by dimension matrix and D is a right singular vector representing document by document matrix [31]. The decomposed matrices are then truncated to a dimension less than the original k-value and the original **X** matrix approximated in the reduced latent space which better represents semantic relationships between terms compared to the original k-dimension document space. A study of LSI indicates its dependency on term-frequency and on the higher level co-occurrence of terms in different documents. Studies also show that LSI ignores the class structure of concepts[56]. The paper presents a technique where LSI is combined with an ontology modelling geospatial concepts that captures class structure and forms links between concepts which may not be captured by LSI alone after the dimension reduction.

3.2.3 Latent Space reduction

Latent Semantic Indexing reduces the vector space by creating a subspace of the matrix dimensions in order to remove noise and redundant terms. The reduced space presents a meaningful association between terms that in turn relate documents[32]. An example of the noise factor can also be seen in figure 3.1 The plot is shown for term-term associations established by the decomposition, where terms like 'in' and 'a' will be close to zero because of appearing next to too many different terms in the original document, therefore a pattern of occurrence is not established. These words are an example of 'noise' and do not carry any meaning with respect to the whole document. In our work, these words have been added on the list of stop words and are removed prior to LSI processing of the document corpus.

3.3 Document-Term Construction using Ontology

The use of an ontology reinforces connections established between terms occurring in separate documents by virtue of a known semantic connection. The ontology serves as a representation of prior clustering knowledge. Using a term-relationship-term triple represented in a knowledge base allows the automation of geospatial web document clustering as it formalises the logic representation of concepts and the semantic links between them. Document clustering is implemented using the proposed method in geospatial web portals in order to group the documents according to the feature observed by the web resource.

3.4 Conclusion

In this chapter an overview of document content extraction is given. The problem is described in the context of the resources published on the web and in particular within the sensor web. An information retrieval algorithm, LSI, is presented and a description of the document and term relationships that are formed by the vector decomposition is given. A definition is provided for the ontology that will be used to bias the LSI. In the following sections we will describe the implementation of the LSI and LSI-ontology.

CHAPTER IV

METHODOLOGY

The work done aims to demonstrate the improvements to the performance of LSI, an information retrieval and document clustering algorithm, applied to large document set for automating clustering and organising the web pages according to page content. The modified algorithm is based on combining LSI with ontology concepts of the topic of interest, for a more accurate document and concept clustering. The original LSI is implemented and evaluated, in comparison with the modified LSI on the same data. The method used aims to show whether: the modified LSI will produce the desired results, and study the performance changes for cluster headings adopted from topics found on the portal. This chapter describes the research method, where the research questions are presented and the formal statement of the hypotheses, and stating the expected results. The research design, which is empirical evaluation of experiments conducted to test the hypothesis is described and a description of the investigation procedure. The experimental data is described followed by the analytical framework, which is the the method and tools used for data gathering and analysis.

4.1 Research question and hypothesis

The research seeks to effectively extract knowledge from html documents (web pages) in order to separate pages that may contain links to sensor resources and irrelevant pages in a geospatial portal. In looking at this, the research attempts to answer these main questions:

- **How to recognise and characterise the relevance of a sensor resource**

  Investigate whether it is possible to efficiently represent a sensor resource in order to guide the content extraction and the extraction of knowledge and to determine whether a page contains information about the feature of interest. How to charactarise a sensor resource and its relation to other web resources.

- **How to guide the theme extraction and bias it to the domain of interest**

  Explore the ways in which a theme can be extracted from a web page to gather semantics that add meaning to keywords in order to be able to predict measured phenomenon on relevant web pages. Does the ontology produce sufficient domain knowledge in order to guide the document clustering, using the topics that are found in the portal in relation to those expressed by the ontology?

- **How the feature of interest is defined and how to store the knowledge**

  How to classify web page content by identifying a feature of interest and store the information for better representation of each portal. What kind of categories are formed? How are pages assigned to the categories? What are the temporal, spatial and organisational characteristics of the stored information? Given the diversity of each web portal, will it be possible to condense the gathered knowledge for the portal?

- **The criteria used for document clustering**

  How best to approximate boundaries in the clusters formed? What structure will be used to store the classified resources? Is it necessary to empirically compute the number of clusters using the k-means algorithm, or can the cluster numbers be assumed from what is known of the portal sub-topics?

- **Performance measures**

  Does the clustering improve significantly using the proposed model compared to blind, unsupervised clustering? Is there a significant computational tradeoff. Modification process to be performed before or after the matrix representation of the corpus? How automated and dynamic is the clustering? Is performance conclusively described by the distance measures, both for inter-cluster and intra-cluster similarity measures?

The expected answers to these questions, in the light of prior research done and related work, are detailed in the hypotheses: It is possible to enhance an algorithm by adding the knowledge. It is expected that the knowledge about a sensor resource can be captured and defined in a rich form for automatic recognition and processing by the information retrieval algorithm. The expressivity of the representation, which will guide the clustering process is expected to have an effect on the performance of the modified algorithm. Increased knowledge is anticipated as the web pages are classified and the structure or form of storing this knowledge could be the same as that of defining and represent a web resource. The corpus size will also have implications on performance. By computing distance measures or cardinal coordinates of the pages withing a cluster, the clustering can be automated, with increased distances from the central theme concept expected to tend to be less reliable.

4.2 Research design

In order to answer the research questions outlined above, an experimental approach is followed in the form of a simulation for the purpose of: analysis, comparison and evaluation of the results produced. The advantage of a validated simulation program is the capability to test and create a working model and have a performance indication before it can be deployed out in the web. Testing of LSI has previously been conducted in similar experimental techniques: of creating a corpus on which to perform the analysis. Therefore a similar environment is created using web pages as the corpus, the web page as the document and the terms on the web page as the objects. The experiment is simulated because the algorithm is not implemented on live web content, but rather on locally stored files that have been fetched from the website. The motivation for this is to create a controlled test environment. In carrying out a simulation it is also possible to perform repeated manipulation in order to determine the effects, without considering the lifespan and accessibility of pages on the web. Fetching a web page individually and following the URLs on each page, in the real web environment would require maintaining a list of unvisited URLs, referred to in [45] as frontier and an additional scheduling algorithm that would determine the order in which the queue of unvisited websites would be navigated. By downloading the website to create a corpus in this work we assume a flat structure and know beforehand how many documents are contained in the corpus, therefore URL queueing is eliminated.

Comparison is between the original and modified LSI algorithm for two sets of portals and evaluation is based on studying the outcome to determine whether the anticipated outcome is the hypothesis stated.

4.3 Methodology

In this section we describe the implementation of the modified LSI algorithm. The issue now is how best to combine an ontology with the LSI processing steps that have been described in Chapter 3. The constituent steps in the LSI process are studied by means of exposing the output of every phase and what the impact would be of introducing modification at each step. The proposed model is then implemented and tests performed. The experiment was carried out to answer some of the questions: Is the proposed model an improvement compared to blind, unsupervised clustering? Is there a significant computational tradeoff? Should the modification process be introduced before or after the matrix representation of the corpus? How automated and dynamic is the clustering?

Clustering using LSI is performed on a single portal, which represents a corpus and then the LSI is combined with ontology concepts and the clustering performed again in order to compare the clustering performance of the two. The process is repeated for the second portal in order to check the consistency of the performance, changes in the clustering pattern and also to check how the portal structure affects the clustering outcome. The clusters of web pages are then automatically classified by using them to populate the knowledge base with instances.

4.3.1 Data Gathering

The tools used in acquiring and storing of the data to be manipulated were as follows: The portal web pages, which are the initial input data, and the output of each of the processing steps are stored locally on a hard disk. GNU Wget, a free Unix tool is used for retrieving the pages to be stored. Download command options which are specified to fetch pages using http are: the wait period between requests, specified to be 1.5 seconds, and the progress indicator which is in the form of a bar, which is useful should a connections timeout occur. The directory structure of the files on the remote server is replicated locally and the retrieval rate is determined by the network download speed. By recursing through a dynamic site, a static version can be generated locally for processing.

4.3.2 Data Processing

Text processing functions were performed using gcc compiler in a linux operating system. C++ Boost libraries were mostly used for regular expressions, which consist of pattern matching functions required for text processing. The boost filesystem is used for directory processing for file manipulation inside the directory structure of the corpus, which is currently not supported by the C++ language. This is chosen because of the flexibility of the libraries across the two platforms used throughout this work: Linux and Windows. The portability allows for us not to worry about, for example, the exact syntax in each platform for separating directories. By specifying a path object in the boost filesystem library, this feature is abstracted. Matlab, which lends itself well to matrix computations, was used for vector representation and calculations. K-means clustering is also used in the Matlab environment, together with the silhouette function, which plots the inter-cluster separations.

Ontology construction is performed using Protege, an ontology OWL editor and knowledge base framework. NaturalOWL, a protege plugin which is a text annotator tool, is used for ontology to natural language conversion. It converts the structured information represented in OWL and converts it to sentences. Automatic instance creation, done by defining an individual in the ontology and to assert properties about them, is implemented in Java.

4.3.3 Experimental data

The aim was to find a document corpus that mainly contained sensor resources, so a geospatial portal were chosen, namely the NASA Global Change Master Directory(GCMD) portal. These contain links to data centres of sensors and sensor platforms. Data requirements of the study were formulated from the need to first identify a sensor resource before being able to discard those that are not. The logic to the proposed approach is that if clusters of smaller distances between connected documents can be formed with relative accuracy, then documents with irrelevant content will result in scattering over larger distances from the formed clusters.

The input data, in the form of html files, used in the experiment was obtained from NASA's Global Change Master Directory (GCMD) portal which is a diretory of earth science data sources that enables both publishing and access to data. The data is published in the form of ancillary descriptions, which is a GCMD structured format for data source, instrument and platform description. The portal links are navigated until the detailed record page is reached. This record signals the maximum depth within the portal along a particular branch and contains the data centre URL which can be accessed outside of the portal website. The record page is in DIF format, a dataset metadata structure employed by NASA. The representation of the sensor resources found in the portal is in the form of searchable fields in a IDN (International Directory Network) database and can be navigated manually by: keyword searches, ancillary descriptions or earth science topic which can be refined by selecting spatial coverage on a map. The GCMD was used as a test portal and also used in the initial manual instance creation and querying to test and compare results to queries entered into the knowledge base versus those returned by the portal.

4.3.4 Procedure

In identifying the tasks to be perfromed, the study was organised into six sequential activities:

1. Simulation environment for testing
2. Html document pre-processing
3. Content retrieval and clustering using LSI
4. Clustering using LSI combined with ontology
5. Euclidean distance comparison as a performance measure
6. Web page classification by inserting pages into the hierarchichal clusters, automatically forming instances of the ontology

There are two aspects of the biased LSI process implementation. The first is developing the ontology which was implemented using the Protg ontology editor in the OWL language development, automatic instance population and natural language document generation using the ontology statements. The second aspect is text processing using html pages as the input, together with the text representation of the ontology as an addition to the original corpus. This involves fetching web pages from the web portal server by invoking wget and arranging the folder creation to follow the same as the website navigation tree. The next step is preprocessing of text, also known as text normalisation. Preprocessing involves applying basic regular expression rules to normalise all characters, tags and expressions appearing on an html encoded script. The overall processing steps are presented, as can be seen in figure 4.1:

4.3.5 Simulation

How best to resemble or approximate the environment where discovery is to take place. An experiment is to be conducted that will demonstrate the results and enable performance evaluation of the system. The conclusions drawn from the system can further be extended in a dynamic environement such as the World Wide Web.

4.3.6 Document pre-processing

Te first step required for term-processing is the the removal of html encodings. HTML tags are elements used to construct the document and determine display properties and contain information about the sections of the document. Removal of html tags includes removing hyperlinks on the page and jpeg tags so that the content is extracted and stored as natural text in a text file. The text file now serves as a string stream input for vector processing. The second step is the removal of stop words and special characters. Articles and conjunctions for example 'and', 'or' are removed from the text so that keywords that will be useful in conveying meaning remain in the text. Different natural language processing tools use different stop word lists and since there is no universal list, a list of stop words used by Google was used as data input for the stop word removal subroutine. The Google list is used by the search engine as a list of words to be ignored when a search is conducted. Stop-word removal also plays a role in scaling down the amount of terms to be indexed and the computational load without affecting the results. The number of words on each web page can be reduced by up to 30 percent due to stop-word removal. The stop word list is modified according to our specific application requirements. For example, some words on the list are removed: eg. fire, which in our application refers to a phenomenon and carries meaning. Some words eg, 'found' which can relate two concepts: such as an entity and its location, are also removed from the stop list. Other words are left on the list because due to the nature of LSI, when the same word appears next to many different terms and a pattern of usage cannot be determined, then it is regarded as noise and will have smaller weighting.

4.3.7 Content Retrieval and Clustering using LSI

Figure 4.1 shows the overall process, with the LSI steps consisting of: Term frequency count, Vector space modelling by representing the corpus as a term-document matrix, Latent space creation by k-dimension reduction. Included in this step is importing matlab vector processing capabilities for matrix manipulation, followed by matrix decomposition and computing a term-term, document-document matrices to establish a pattern of occurrence and finally perform document clustering.

4.3.8 Content Retrieval and Clustering using LSI combined with Ontology

A domain specific ontology is combined with LSI by including the ontology concepts in the corpus. Ontology structures are converted to natural language, as an ontology pre-processing step, and included in the corpus after html to text parsing has been

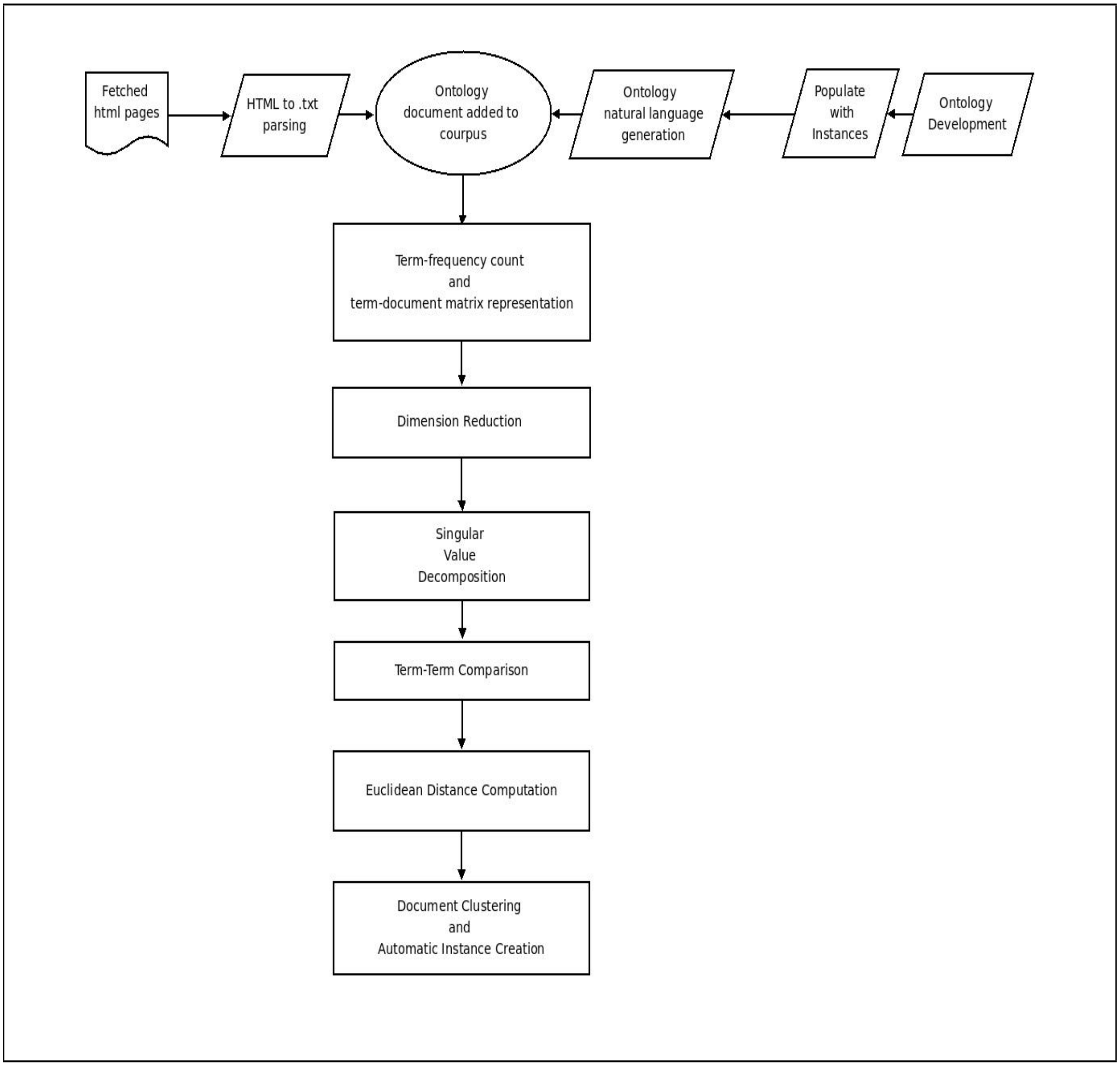

Fetched html pages
HTML to .txt parsing
Ontology document added to courpus
Ontology natural language generation
Populate with Instances
Ontology Development
Term-frequency count and term-document matrix representation
Dimension Reduction
Singular Value Decomposition
Term-Term Comparison
Euclidean Distance Computation
Document Clustering and Automatic Instance Creation

Figure 4.1: Flow of the modified LSI process

perfromed.

4.3.9 Construction of the matrix using MATLAB

A textscan function is performed on each of the corpus pages. Each page is scanned for unique terms that are indexed and against the frequency of occurrence and stored in a cell array to form the local term frequencies for each page. A unique term is an individual word that is extracted from the body or a portion of the body, for example if it is only headings to be used to capture the page content. These local frequencies are combined to form the global frequencies for the corpus by concatenating the frequencies stored for each term appearance on each document. The global frequencies form the term-document matrix shown on table 4.1, with the corresponding occurrence for each term on the list for each document. The frequencies determine the weight of the term which reflects the importance of the term for each document and further used to compute the relative importance throughout the documents. The initial term-document matrix has a document-term dimension of: 22568 x 2881. From this matrix the SVD is performed in order to determine the document-document clustering.

4.3.10 Euclidean distance computation

Euclidean distance is used as the measure for cluster formation, according to central themes, that the documents are nearest to. The details of the distance evaluation metric are presented in the results section of Chapter 5.

4.3.11 Cluster Analysis

K-means clustering is performed on the document set. Using the Euclidean distances to determine cluster membership. The clusters formed using the modified LSI combined with ontology are further sub-clustered using the linkage function to form hierarchical sub-clustering.

4.3.12 Creating Instances for web page classification

The final step is automatic instance creation, which is automating the realisation process of slotting a WRSR instance under a relevant class based on document content. This automatic classification of web resident sensor resources found in portals contributes in the storage of clustered pages into a class, therefore serving as a more

Table 4.1: Term frequency matrix corpus terms subset

	Documents Indexes: Web pages containing the terms					
Terms	doc(1)	doc(2)	doc(3)	doc(4)	···	doc(22658)
agriculture	4	2	2	2	···	1
altitude	2	2	2	2	···	1
atmosphere	8	2	2	12	···	1
behaviour	2	4	2	2	···	1
biological	2	4	40	2	···	1
biosphere	2	8	2	2	···	1
canada	2	4	2	2	···	1
caption	10	2	4	2	···	1
citation	2	2	2	2	···	1
city	2	4	4	2	···	1
classification	4	2	2	2	···	1
climate	2	2	4	4	···	1
collection	2	2	4	12	···	3
commission	2	2	2	2	···	1
coordinates	6	2	6	2	···	1
country	2	2	2	2	···	1
coverage	4	2	2	2	···	1
cryosphere	2	2	2	2	···	1
dataset	28	2	2	2	···	1
date	12	2	2	2	···	7
description	2	2	12	2	···	2
digital	4	2	2	24	···	1
dimensions	2	2	14	4	···	1
document	6	2	2	4	···	1
downloadable	2	2	2	2	···	1
earth	2	2	2	2	···	1
environment	2	2	2	4	···	2
feature	2	2	2	2	···	1
···	···	···	···	···	···	···

meaningful representation of the portals.

4.3.13 Analysis

The last step is analysing the results in order to evaluate the effectiveness of the proposed modification to LSI. By tabulating the distances between the central theme and documents of similar and relevant content we compare the implementation of LSI on its own and that combined with an ontology. Finally we show an improvement to document retrieval and discovery within the constructed knowledge base than with keyword searches in both portals. The limitations of the work done are also considered, for example the performance of LSI with respect to the size of the corpus and also the computation complexity of matrix representation for very large document-term dimensions.

4.4 Conclusion

The method followed for implementing LSI combined with ontology is presented. The steps are described for LSI simulation using a corpus consisting of geospatial portal pages, testing of the performance by computing Euclidean distances for the clusters formed and compared to that of LSI on its own. Similarity is measured within each cluster, where minimum distances indicate high document similarities. The second performance indicator used is inter-cluster distances, where the higher the distances of separation of the clusters, the higher the dissimilarity of the documents placed in different clusters. The clusters formed by the modified LSI-ontology algorithm are used as the document input in order to form class hierarchies within each cluster. Finally a description is given of automatic knowledge base population for storing the information acquired in the information retrieval process. Chapter 5 gives a detailed description and analysis of the results and the discussion of the overall performance of the modified LSI-ontology in comparison to the original LSI.

CHAPTER V

ANALYSIS AND RESULTS

A document classification model is implemented in order to enable resource discovery for a geospatial portal, the NASA GCMD portal for Earth Science Data and Services. A process is implemented that classifies 22 658 web pages for the web portal that forms the document corpus. Information Retrieval is performed on the original corpus using LSI, a vector model algorithm as described in Chapter 3. This is compared to a modified implementation of LSI, which makes use of an ontology. The modified algorithm is referred to as LSI-ontology(LSIO). The purpose of the ontology is to define a web sensor resource and store instances of the resources. The ontology is created and then represented in natural language to be included as a text document that forms part of the corpus. LSI is then implemented on this modified corpus and its effect on clustering is compared with that of the original LSI implementation. The aim of including the ontology is to bias the classification to that of resources of interest to prove that prior knowledge domain can be used to enhance the performance of LSI.

In this chapter the results are presented and analysed to show the improvement in document clustering using the two methods. Performance of the modified LSI, and LSI-ontology, is measured by comparing the accuracy with which a relevant document is assigned to the correct cluster. The accuracy is determined by comparing the average distance between the document and the cluster centroids of the clusters for each of the two LSI implementations. The analysis is presented as follows: first a description of the original document groupings as arranged in the portal is presented. The second section describes the method used for determining the reduced latent space, followed by cluster formation and analysis. Finally the discussion of the results is presented, followed by the conclusion.

5.1 Portal structure

The portal structure is analysed for a comparison to be made with the clustering categories formed by LSI. The portal pages are arranged according to the measured feature, which lends itself well to the WRSR ontology feature subclasses. There are three categories for portal navigation: Parameters, Data Centres and a combination of Keyword and Spatial search. The parameters are organised according to the feature being measured, for example drought and precipitation indices which are grouped air temperature and humidity indices under the category: climate indicator. A tree structure with a depth of 3 levels, is used for grouping measurements of the same features, eg. drought and precipitation are further broken down into crop and fire indices. The second option for performing a search is a Metadata search which is performed by navigating through the resource description categories eg. Data centre, project name and Instrument or platform descriptions. The third option is to perform a full text search which uses keywords combined by boolean operators to retrieve information.

The directory structure of the original site is ignored and a flat structure is recreated to form the corpus, after fetching the web pages. The recursive file retrieval options are set such that the links on each page are followed but referenced home pages outside of the the current GCMD domain server are ignored by not allowing the spanning of hosts.

A comparison is made of the current portal structure and of the clusters formed using LSI. This is done by extracting information from all the web pages found in the portal, from all categories (as they exist) to form the corpus. LSI will restructure the portal pages and grouping of related resources. A section of the portal is chosen as the resources of interests to be discovered and to test the performance of the LSI versus the LSI and Ontology modification. The performance of the LSI-ontology is measured by computing the distance of the clusters formed. The focus of the ontology is on water-relate resources. The water related resources are those classified under: oceans, climate indicators, cyrosphere and terrestrial hydrosphere in the portal. The rest of the resources are classified under: agriculture, biosphere, solid earth, etc.

It will be shown that the guided LSI-ontology, performs better clustering and separating of the resources of interest, therefore, enabling discovery of water resources, specifically solid water or ice, within the portal on a better scale than just using LSI for the classification and retrieval of the web documents. The solid water related resources are then stored as instances of the ontology, which captures the knowledge gained from clustering the web pages found in the portal.

5.2 Dimension Reduction

The results are presented to compare the performance of Latent Semantic Indexing(LSI) and LSI-ontology(LSIO). A term-matrix of dimension [22658x 2881] is generated which is decomposed. Dimension reduction, also referred to as latent space representation, is performed on the decomposed document, term and singular vectors by approximating the original matrix using the k largest singular values. The truncated vectors in the reduced space can be multiplied out to approximate the original matrix. Determining the k-value is crucial to the performance of LSI for removing the noise (also referred to as a sampling error) caused by terms with large co-occurrence, affecting the accuracy of the term-document clustering. Reduced dimension also has the additional advantage of reducing computational complexity.

In this application the aim is to reduce the Singular matrix from [22658 × 2881] to [22658 × k], with k << 22658. A large k-value may result in noise and an over-reduced k-value may mean the loss of important data and meaning. It is suggested in the original implementation of LSI by Deerwester that this value be computed empirically as it is affected by the content and the size of the corpus[15]. Different documents sets have different pattern of usage of terms and typically the value of k should be less than 300. Work has been done and various methods have been proposed with regard to finding the optimal k-value in LSI dimension reduction. One approach, proposed by Everitt is to find k by calculating the relative variance of the Singular vector values[18]. The variance is calculated using the equation:

$$V_i = \frac{S_i^2}{\sum_{j=1}^{r} S_j^2}; i = 1, 2, 3....r \tag{5.1}$$

where, V_i is the variance for each singular value S_i relative to the square sum of all r singular values. The variance is calculated for a maximum of $r = 300$ singular values. Figure 5.1 shows the variance values that are used to determine the k most significant values to be used for the reconstruction of the term-matrix. The Everitt method states that the first k-values which are greater than $0.7/n$ should be selected, where n is the total number of terms of relevance. A combination of the variance method and empirical computation is used to determine the k-value, which for the GCMD portal term-matrix is determined to be 10. The variance plot shows a sharp decrease for $k < 10$, and then tends to approach a constant. The cutoff is chosen for k=10 with a variance of 0.0062672, after which the values of $0.7/n$ drop significantly from 0.0054894 to 0.00036452, where n=2881. A hybrid of both methods: variance computation as shown in the variance plot and the Everett method of the first k values of singular values greater than $0.7/n$ is used to arrive at k=10.

The plot shows that a constant variance is approached for values $k > 10$ and a rapid decrease for $k < 10$.

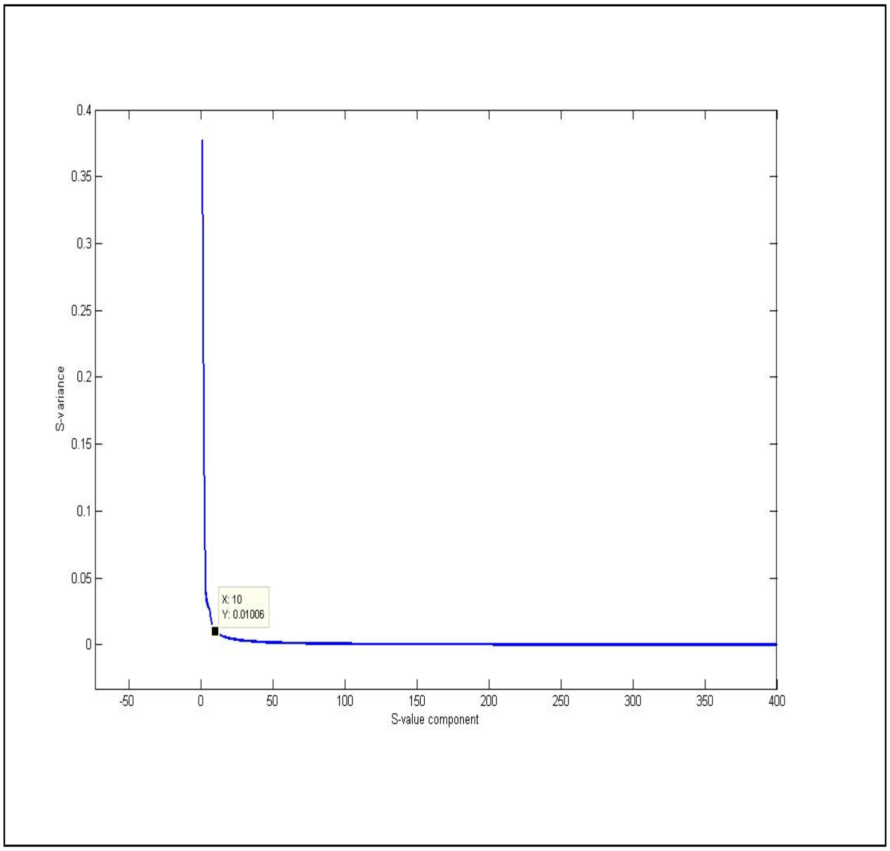


Figure 5.1: Relative Variance of singular vectors

5.3 Cluster Analysis

The document-document vector is represented in the reduced semantic space as 10-dimensional coordinates in space. K-means clustering is performed on the data by creating K mutually exclusive clusters of data, in this case groupings of the 10-dimensional co-ordinates and K centroid coordinates. The objective of the K-means algorithm is to partition the data by minimising the sum of distances between each data point and the centre of the cluster, the centroid. The distance measure used is the Euclidean distance measure, in Euclidean n-space defined by the equation:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \vdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (5.2)$$

where distance is calculated between two points P and Q, each with n-dimensional co-ordinates. K-means clustering uses residual sum of squares (RSS) as the objective function, which is to iteratively minimise the squared distance of the cluster members from the centroids. This is a measure of how well the centroid represents its cluster members (document set). The centroids are intially randomly selected and for each seed value square sum of distances is calculated for the surrounding data points. The number of clusters to be computed is specified as the input k. The cluster centroids are moved around until the criterion of minimum squared distance is met. Choosing the correct value of k is an important issue, addressed widely in literature[53]. It is suggested by Manning that this can be determined by performing a heuristic search, or can be based on what is known about the domain[37]. In the case of the GCMD portal, it is known that the documents occur in 14 categories, as listed in table 5.1, therefore this is the optimal k that will be input. Intra-cluster similarity is computed for a range of k inputs in order track the performance of both LSI and LSIO and the intra-cluster distance used to investigate which is of superior performance.

Table 5.1: Cluster topics of GCMD documents

Cluster	GCMD Topic	Portal Composition(%)
1	Agriculture	4.31
2	Atmosphere	14.95
3	Biosphere	7.96
4	Biological Classification	13.33
5	Climate Indicators	0.88
6	Cyrosphere	4.94
7	Human Dimensions	8.42
8	Land Surface	10.30
9	Oceans	12.36
10	Paleoclimate	3.31
11	Solid Earth	6.25
12	Spectral Engineering	4.85
13	Sun Earth Interactions	0.84
14	Terrestrial Hydrosphere	7.30

Low intra-cluster average distances translate to a high similarity measure between the homogeneous cluster elements.

It is difficult to visualise the clusters as a result of the $n = 10$ dimension, therefore emphasis is on cluster analysis in order to compare LSI and LSIO as follows:

- **Cosine similarity between any two documents**

Similarity measures used are that of the Euclidean distance between any two documents and their respective angle of similarity. A Euclidean distance matrix for each document with respect to all the other documents is presented, calculated using the 10-dimension co-ordinates of each of each document obtained from the LSI document-document output, where any two vectors define two points in a 10-dimensional space. The Euclidean distance is the absolute distance between these two points. The higher the similarity between two documents, the smaller the Euclidean distance. For each of the distances and the coordinates we can compute an angle between any two documents in a 10-dimensional space. This is the angle of similarity, where the cosine of the angle lies in the range $[-1, +1]$. The similarity is measured by how close the value approaches 1, where $+1$ represents maximum similarity (any document with itself) and -1 represents maximum dissimilarity. Dissimilarity is when two document vectors are facing opposite directions. The combination of two similarity measures is used to represent the clustering of the documents.

- **Intra-cluster average similarity**

Evaluating the clustering by analysing the documents within one cluster. The goal is to have minimum average distances within each cluster, which is an indication of high intra-cluster similarity. The intra-cluster average distance is determined using the equation:

$$\mathbf{D}_{ave}(k) = \frac{1}{N(k)}\sum_{i=1}^{N} \mid \vec{x}(i) - \vec{\mu}(k) \mid^2. \tag{5.3}$$

given the centroid μ of the k-th cluster with N cluster elements represented by spatial location $\vec{x}(i)$

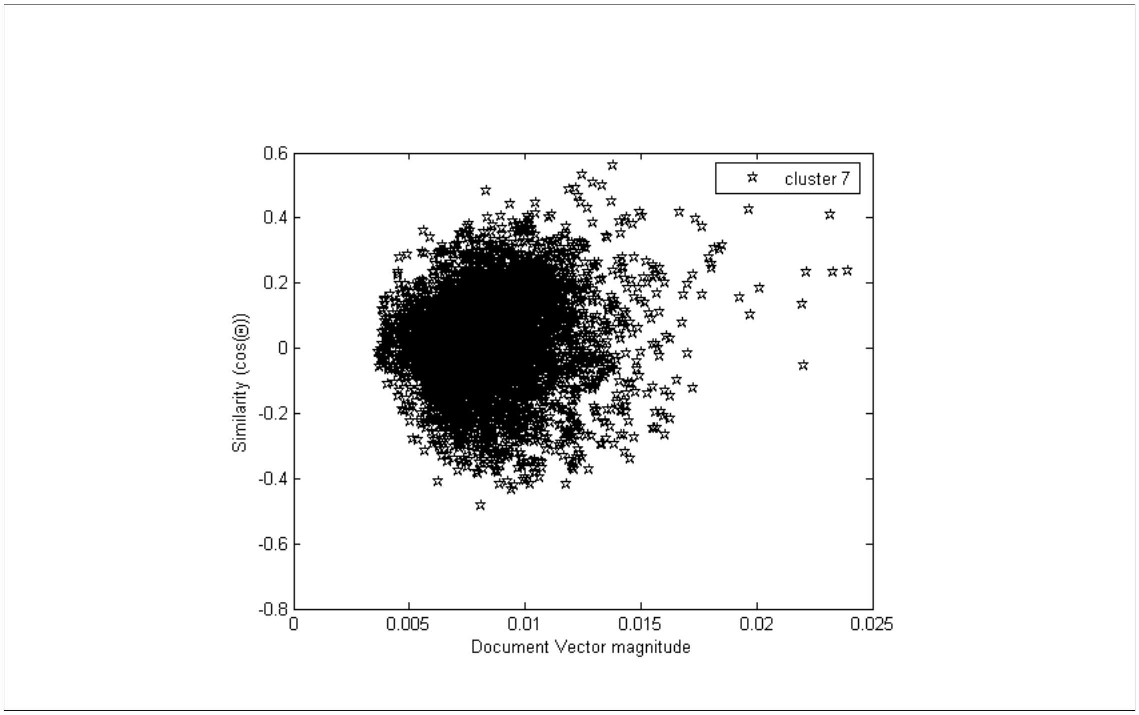
- **Inter cluster separation**

The distance that separates two clusters indicates the dissimilarity between the clusters. In flat clustering the objective is to have similar documents belonging to the same cluster, and dissimilar objects in different clusters and far away from each other. A silhouette plot is used to indicate how well separated clusters are from each other. For each point in the cluster, a silhouette value is calculated, indicating how far it is from points in other clusters. The silhouette value interval is [0,±1], where 1 is maximum separation. Silhouette plots are also useful in evaluation of clustering, as an indication of the quality of the clustering.

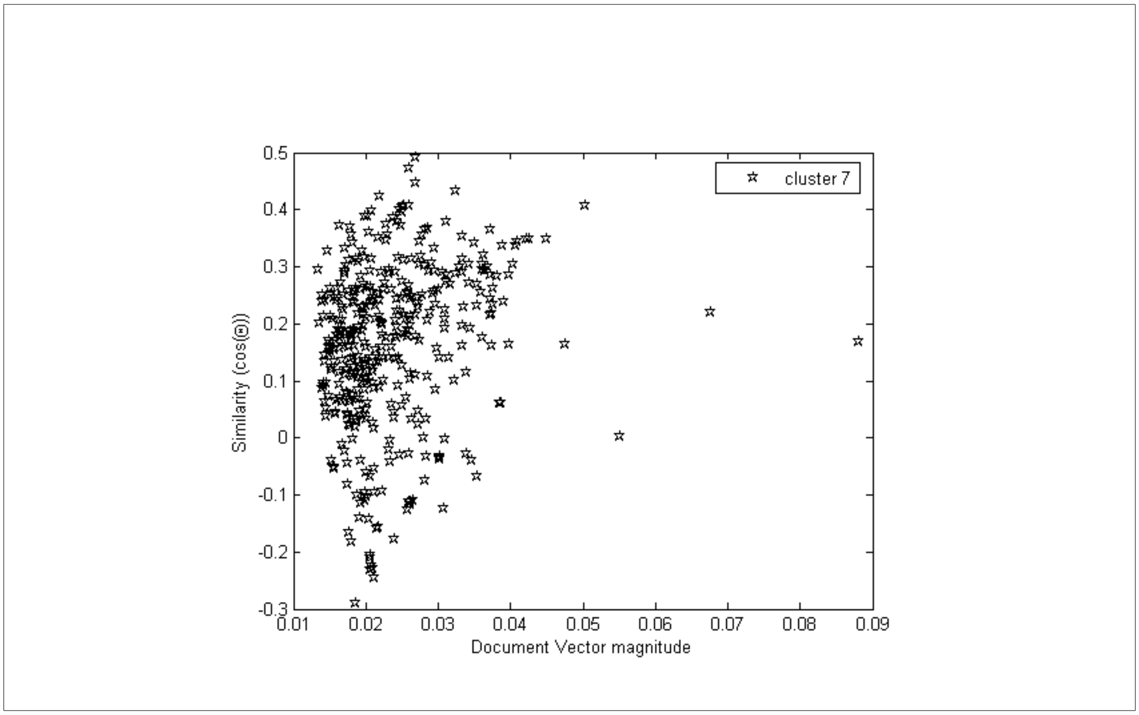
5.3.1 Cosine Similarity

In figure 5.2(a) and 5.2(b) we see the plot of Euclidean distance between each document plotted against the similarity between two documents. The clustering plot is only to show the significant difference in cluster membership for LSI and LSIO for cluster 7. The true spatial relationships of the clusters cannot be visualised in 10-dimension. The cluster analysis is represented and explained for all clusters using the silhouette and average distance computation.

Table 5.1 shows the topics found in the GCMD portal and the percentage composition of documents in each category. The largest document sets belong to: Atmosphere, Land Surface and Atmosphere. From these categories the water related documents of interest will be clustered together and inserted as instances in the water ontology that has been created. The aim is also to create exclusive membership of documents, where no document occurs in more than one cluster. The way the documents are arranged in the portal is such that a document occurring under the



(a) LSI



(b) LSI-ontology

Figure 5.2: Vector magnitude vs cosine similarity for cluster 7

category 'Atmosphere' may also be found under 'Sun Earth Interactions' therefore the portal composition is not a true representation of the document groupings found.

5.3.2 Intra-cluster average similarity

The results for intra-cluster average similarity are presented. Increasing k-values are selected as input, from k=2 to k=15, where the optimum k is stipulated as 14, as the known number of categories for the GCMD portal documents. The aim of computing different k-clusters is to compare the average intra-cluster similarity for each. The goal is to show that regardless of the input value of k, lower intra-cluster distances are achieved by LSI-ontology. Therefore the focus lies in the arrangement of the documents after including the ontology, which will make the LSI-ontology cluster better regardless of the clustering mechanism used. Taking this into consideration, we focus less on the accuracy of the k which is chosen.

Figures 5.3 to 5.6 show plots for LSI and LSIO on the same axis for increasing values of k. It can be seen that the LSIO plot generally have low distances, compared to the LSI. As the number of cluster increases, the more evident this is. In figure 5.3, where there are four clusters we see that LSIO has higher average distances for cluster 2 and cluster 3. For higher partitions, where 8 or 9 clusters are computed we see that in most clusters the LSIO has lower average distances. For k=14, we see that the two plots have almost equal average distances, where for cluster 9, 10 and 11 we see that LSIO has slightly higher average distance. When the number of clusters is increased further to exceed the 14 clusters, it is seen on figure 5.6 that once more the LSIO has superior quality of clustering, by looking at the lower average distances. From cluster 6 to cluster 12 of the k=15 plot, there is a noticeable difference in the distances attained for LSI and LSIO.

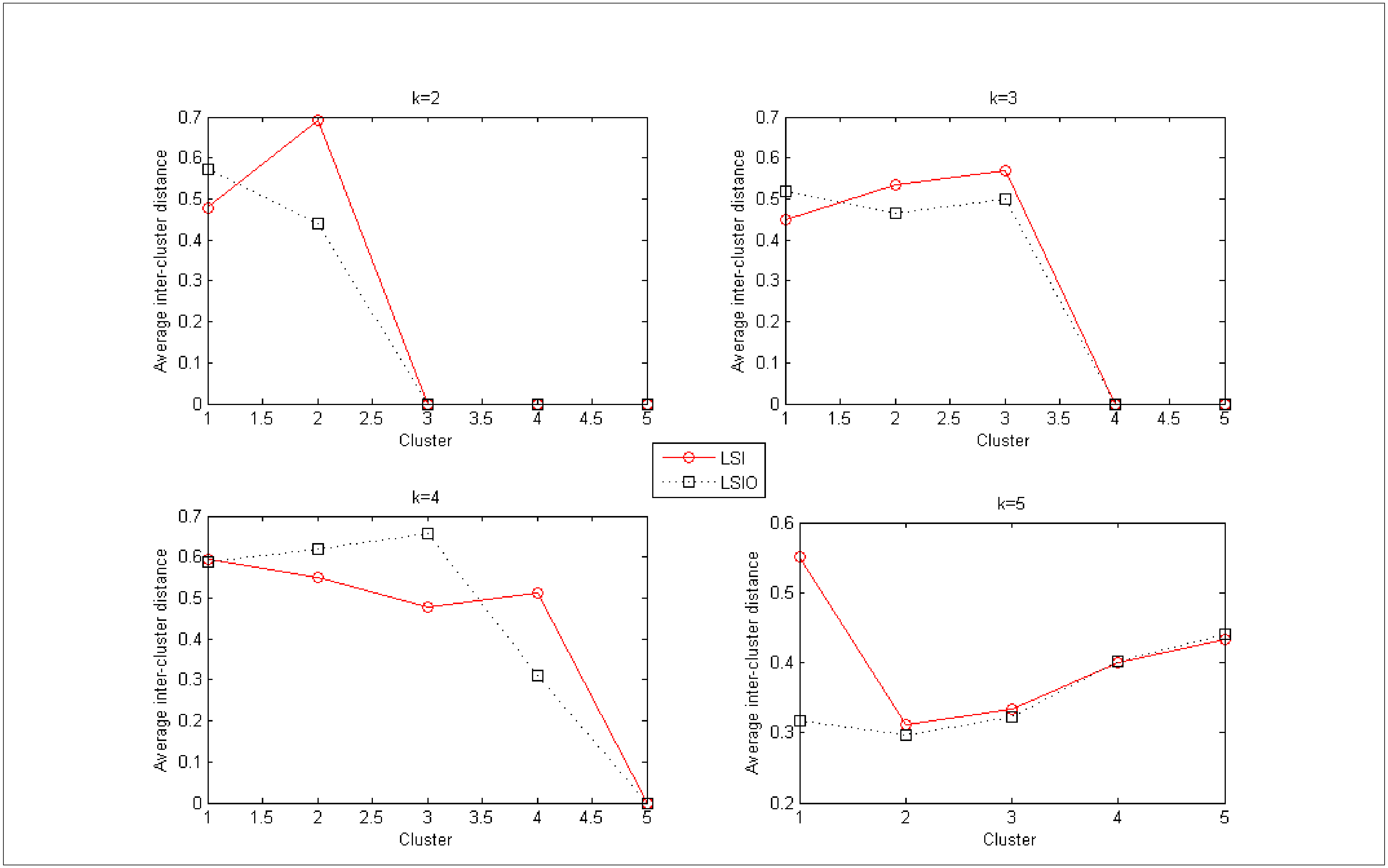


Figure 5.3: Intra-cluster average similarity, k=2,3,4,5

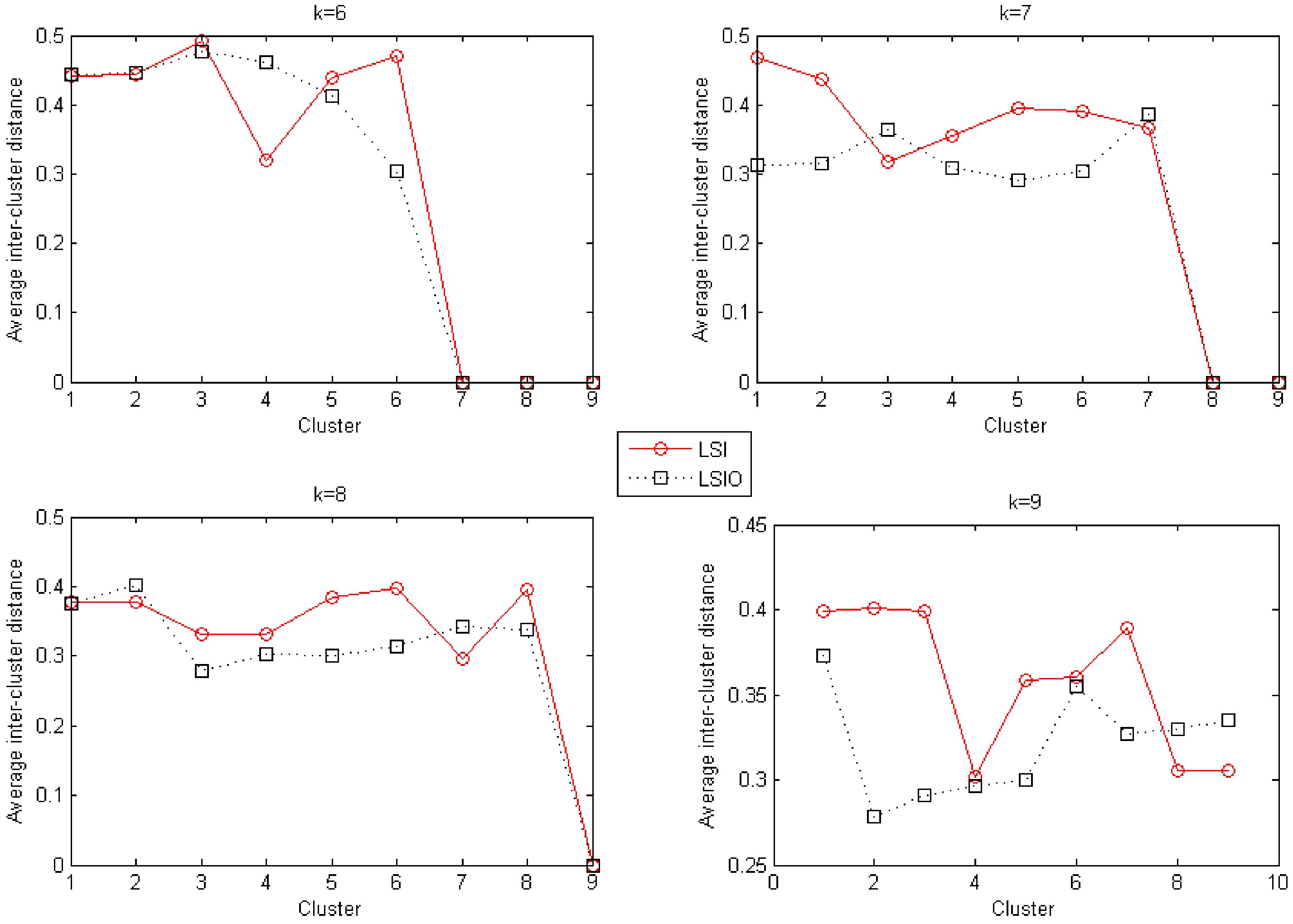


Figure 5.4: Intra-cluster average similarity, k=6,7,8,9

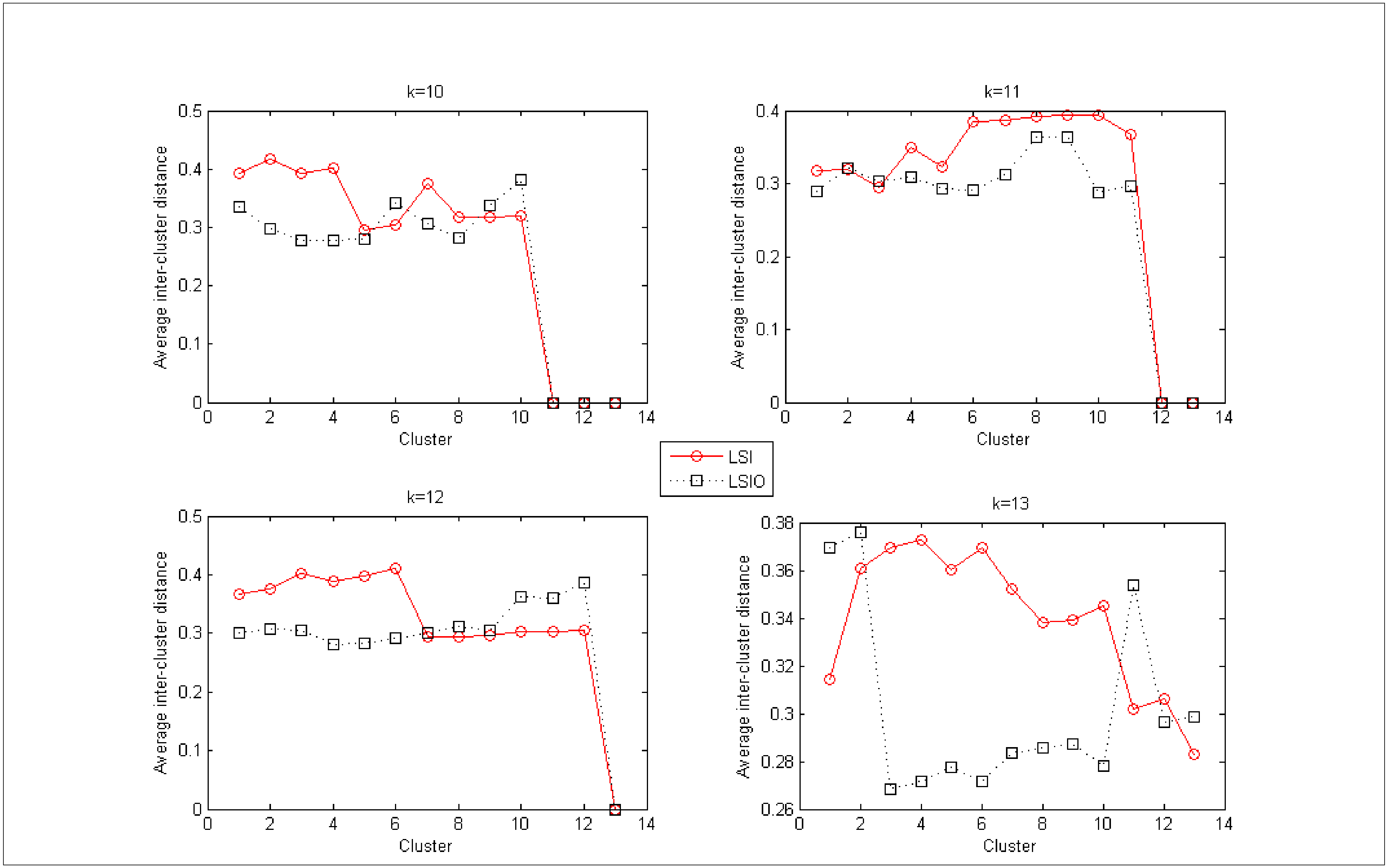


Figure 5.5: Intra-cluster average similarity, k=10,11,12,13

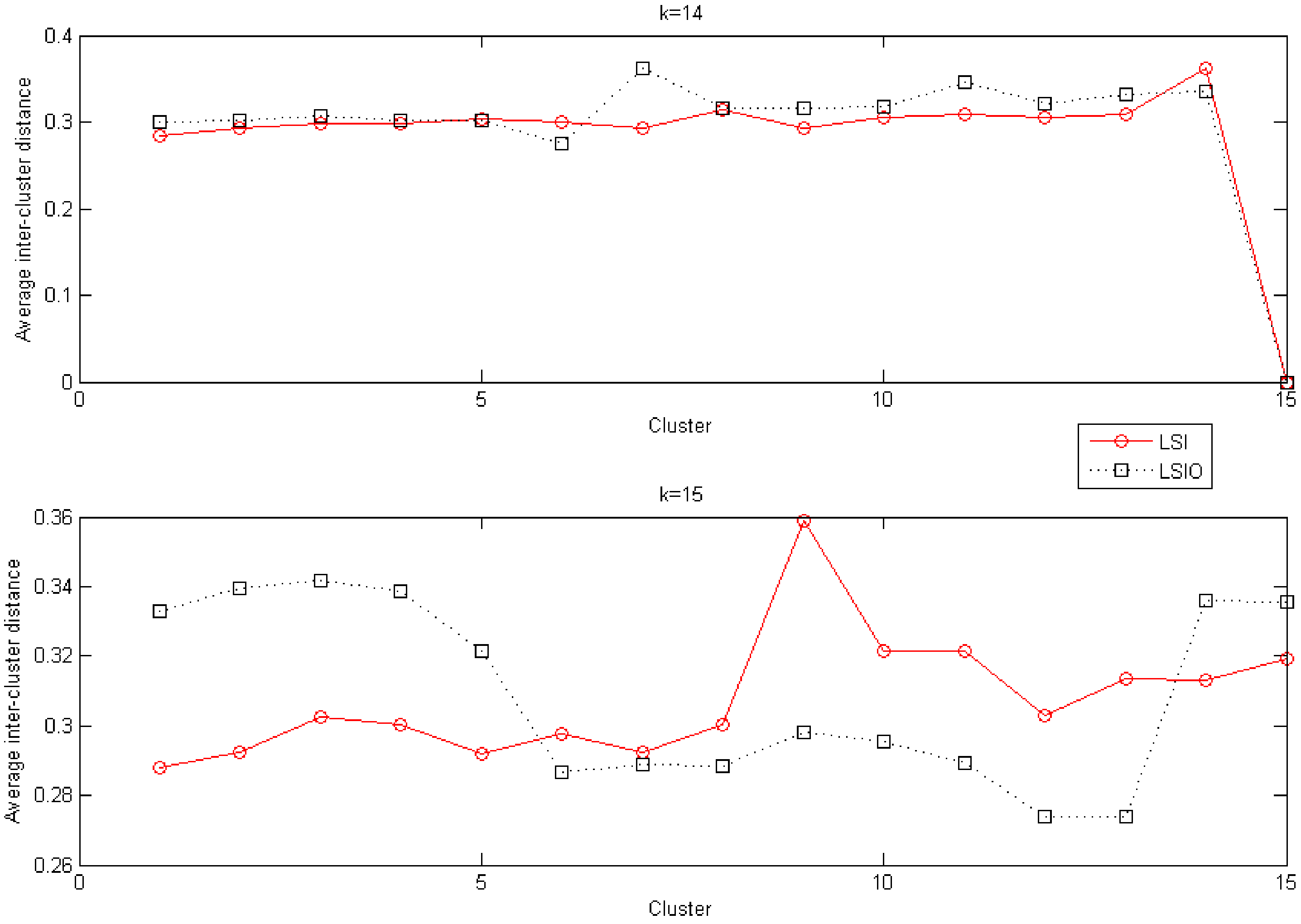


Figure 5.6: Intra-cluster average similarity, k=14,15

5.3.3 Inter cluster separation

The similarity measure alone is not enough to conclude the performance of LSI versus LSIO clustering quality. We further look at inter-cluster separations. Silhouette values are computed to show how dissimilar each document within each cluster is to neighbouring documents in other clusters. Each document in a cluster has a silhouette value computed which is in the range [-1,+1]. If the silhouette value plotted for a document is 0, this means that it lies very close to a document from another cluster and 1 means that the document lies far away from any document of any other cluster. The silhouette plot therefore shows the cluster separation for each cluster, and it can be seen just from the shape of the silhouette if a large number of objects in that cluster lie close to other cluster members. The silhouette plot also shows cluster membership, where a large silhouette area indicates a higher number of documents that have been classified under that particular cluster. From the plots in figure 5.7 and 5.8 we see the silhouette plots of LSI and LSIO respectively. The separations are expressed as a Euclidean distance measure. Each of the clusters, numbered 1 to 14 correspond to the 14 cluster themes which as indicated previously.

In figure 5.7 it can be seen that the cluster which is well separated from the rest is cluster 8. The silhouette values reach a maximum of 0.9. It is also the cluster with the most documents. The next two clusters with the most documents is cluster 2 and cluster 11. Cluster 5 has the least number of documents and silhouette value of less than 0.2 indicate that the documents in this cluster are not well separated from the rest. They lie close to the partition of a neighbouring cluster.

Figure 5.8 shows the silhouette plot for LSIO. Cluster 13 is the most well separated, with maximum silhouette value equal to 1. Cluster 5 and cluster 7 have the highest document membership. From the silhouette plots it can be seen that the documents

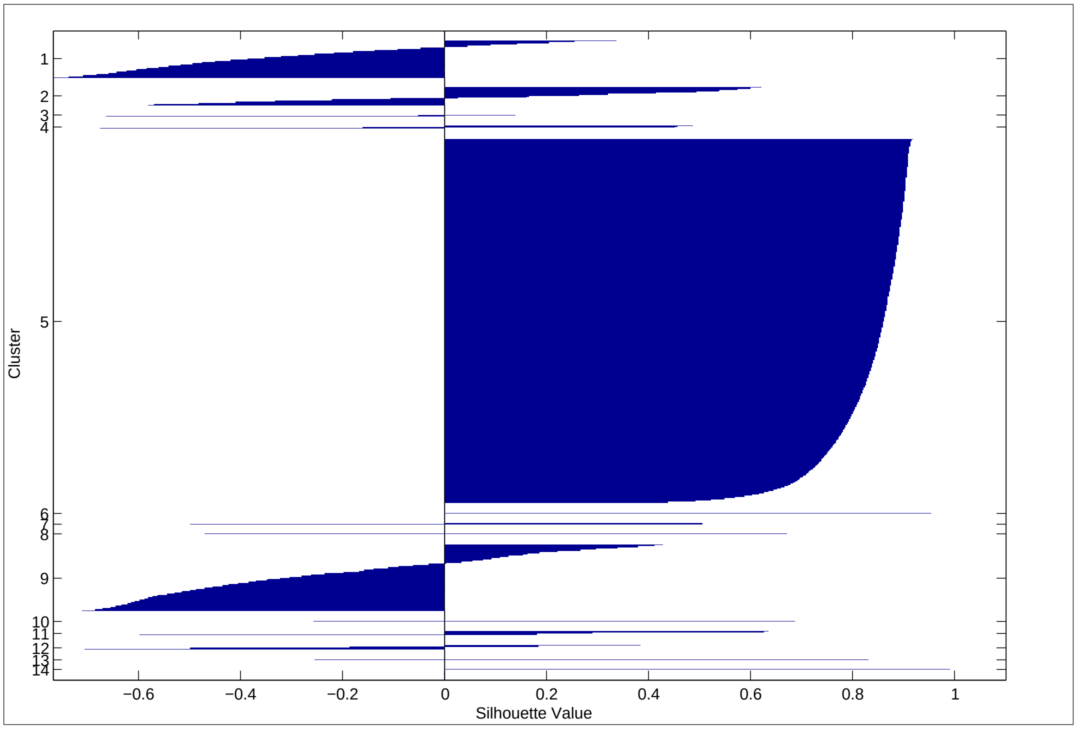


Figure 5.7: Inter-cluster squared distance separation (LSI)

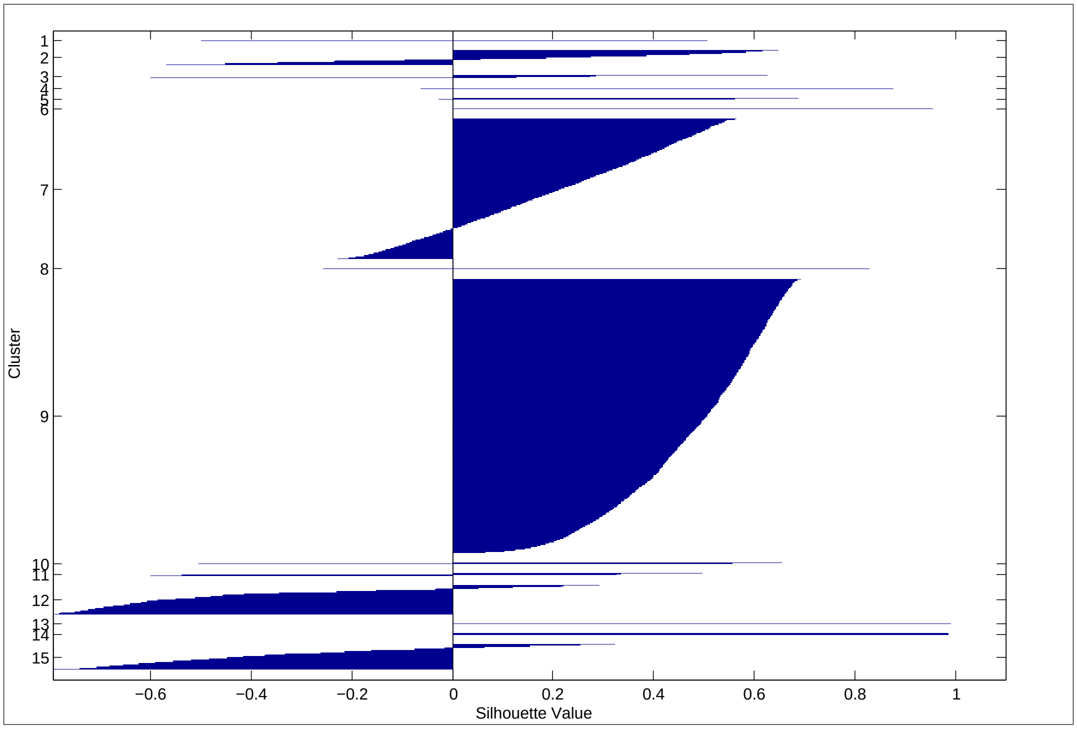


Figure 5.8: Inter-cluster squared distance separation (LSI-ontology)

are more spread out compared to LSI clusters, where the documents were all classified in cluster 8. The document set represented by LSIO shows that there is an additional group of documents that lie outside the 14 clusters that are formed. Cluster 15 shows the documents which are outliers, which means that they lie far from any of the documents and do not fit well into any cluster. It is expected that some documents in the portal do not contain any relevant information, regarding sensor data sources. The relevant pages in the portal are those that contain information about the records and a description of the platforms used to record measurements. Examples of these non-relevant pages are the home page and media update pages such as a list of upcoming NASA conferences. These pages do not contain direct reference to observed features contained in the ontology and are therefore grouped separately, as can be seen in figure 5.8.

5.4 Instance Creation for Web Resident Sensor Resource ontology

It is seen from the intra-similarity plots that the average clusters within each cluster are minimised by the inclusion of the ontology in the corpus before LSI is performed. Using the ontology clusters, we further form sub-clusters using hierarchical clustering in order to form sub-groups of similar objects within each theme. The cluster tree is formed in order to create groups of similar documents that will be used to populate the knowledge base as instances belonging to a certain subclass. An illustration of this process is performed for the two largest clusters, but can be implemented for all 14 clusters in order to further categorise every document found in the portal. The two largest clusters formed are cluster 5 and cluster 7. The number of branches of the tree are specified as an input. The depth of the cluster tree can be increased for further classification by increasing the number of leaf nodes, therefore forming more branches of the tree.

The linkage cluster trees seen in figures 5.9 and 5.10 show the dendrogram plots formed by the linkage function for cluster 5 and cluster 7. The height of the vertical lines connecting the nodes of the clustering represents the distance between the two groups of objects being connected. Distance represents similarity of themes, with the greatest distance indicated unrelated topics. The output of the dendrogram plot returns a vector with elements corresponding to the documents categorised under each leaf node. The document node groupings are used to locate the terms that occur in each of the documents, and these terms are used to compute the theme of each node. The relationship between terms and documents is represented by term-term vector space locations that correspond to the document-document locations in each cluster. In figure 5.9 it can be seen that nodes [8,2,1,7] relate to documents with themes: Antarctica and glaciology and nodes [3,5,4,9,6] have themes: instruments, surfaces and regions. From the diagram we also see that node 10, which has one level or depth, has themes: nutrition, march, dates. The relationship between the themes of node 10 cannot be automatically extracted, therefore in the case of creating instances for the ontology it would require additional sub-classing in order to further separate documents from the parent branch.

Figure 5.10 shows that for nodes [1,2,3,5] themes are: hydroxyl, composition, chemical. The themes found in these nodes fall under the parent node with themes: wind, signals and volcanic. Similarly the themes for nodes [8,9]: radioanalytical, samples, uncalibrated. The number of documents in each node also indicate that further classification can be performed in each sub-cluster until the documents can all be classified under the classes of the ontology. The larger the percentage document membership in the cluster, the more terms that occur with different themes. Sub-clustering creates smaller groupings with more specific themes.

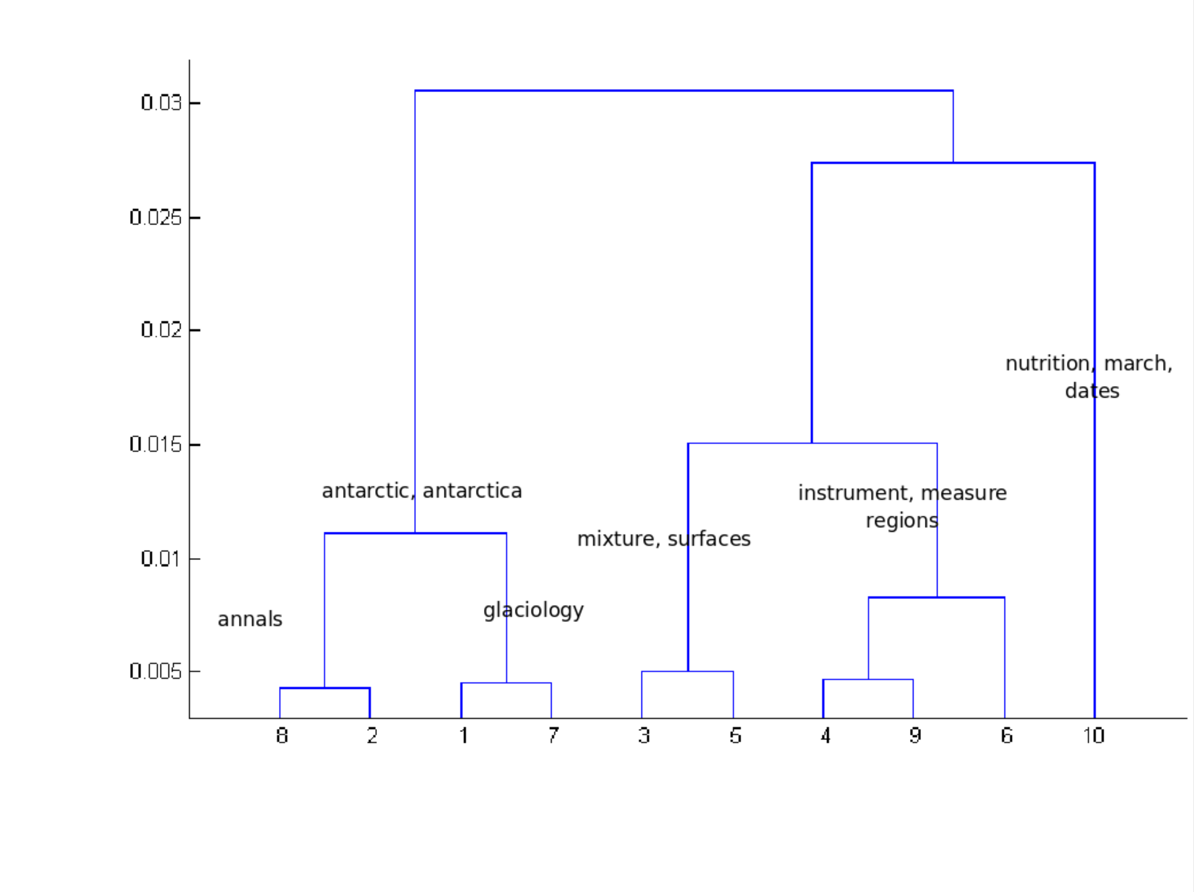


Figure 5.9: Hierarchichal cluster tree for cluster 5

The hierarchy of clusters formed in figures 5.9 and 5.10 based on Euclidean similarity calculations is used to align the clusters with the sub-classes found in the water ontology that was used to modify the LSI theme extraction. The ontology has existing sub-classes of the class WaterBody, as seen in 5.11 and is further broken down into LiquidWater, SolidWater and WaterVapour. From the sub-clusters formed by the linkage calculations we are able to see two groups that can readily be aligned to the ontology. Instances are created in the ontology by slotting the documents into the corresponding classes. The description of the class sensor, which is defined in the ontology as a physical object for measuring a feature, fits with the cluster 5 nodes [3,5,4,9,6].

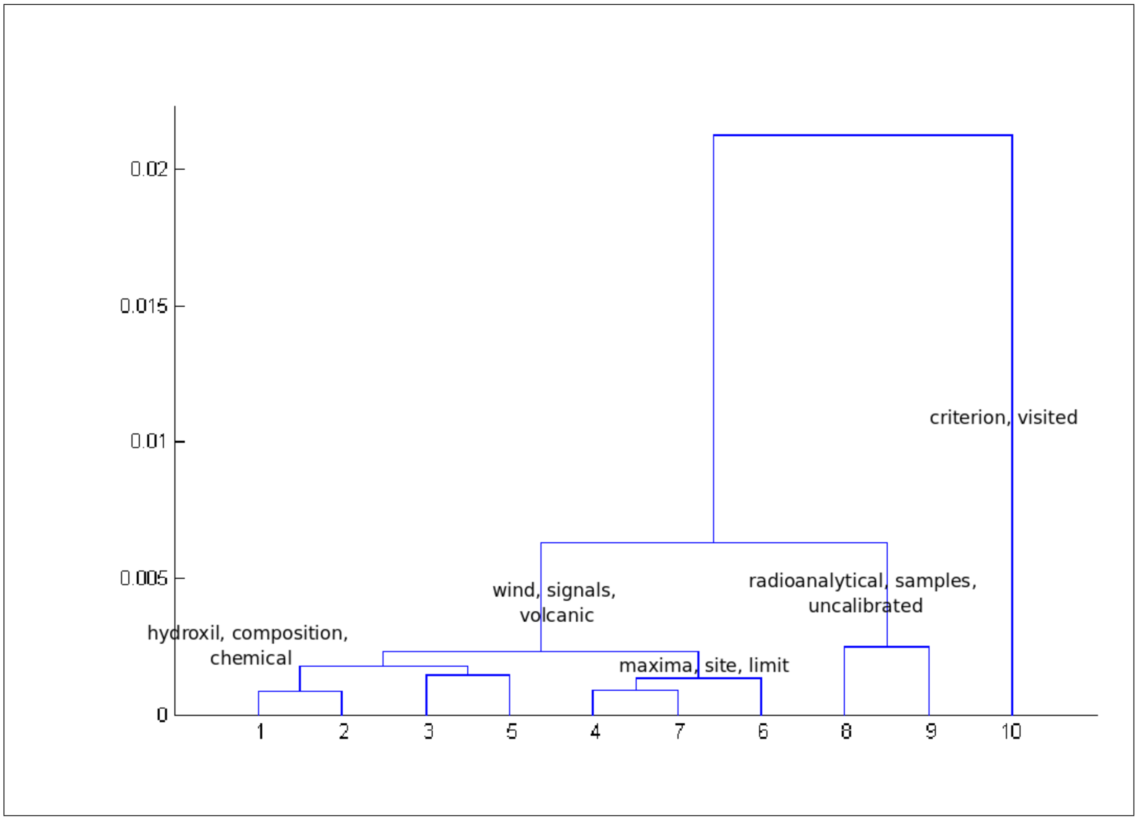


Figure 5.10: Hierarchichal cluster tree for cluster 7

These nodes indicate that the group of documents contained has themes around surface measurements and instruments. This class can be broken down further by cub-clustering in order to group the specific instruments and sensors that are described by each document set. A second case of instance creation is performed by slotting the documents found in nodes [8,2,1,7] of cluster 5 which are based on documents containing snow and ice data, with the themes: glaciology and location : Antarctica. This is an example of the meaning added by the groupings formed using LSI-ontology where in LSI the association of arctic and solid water is solely dependant on term-occurrence.

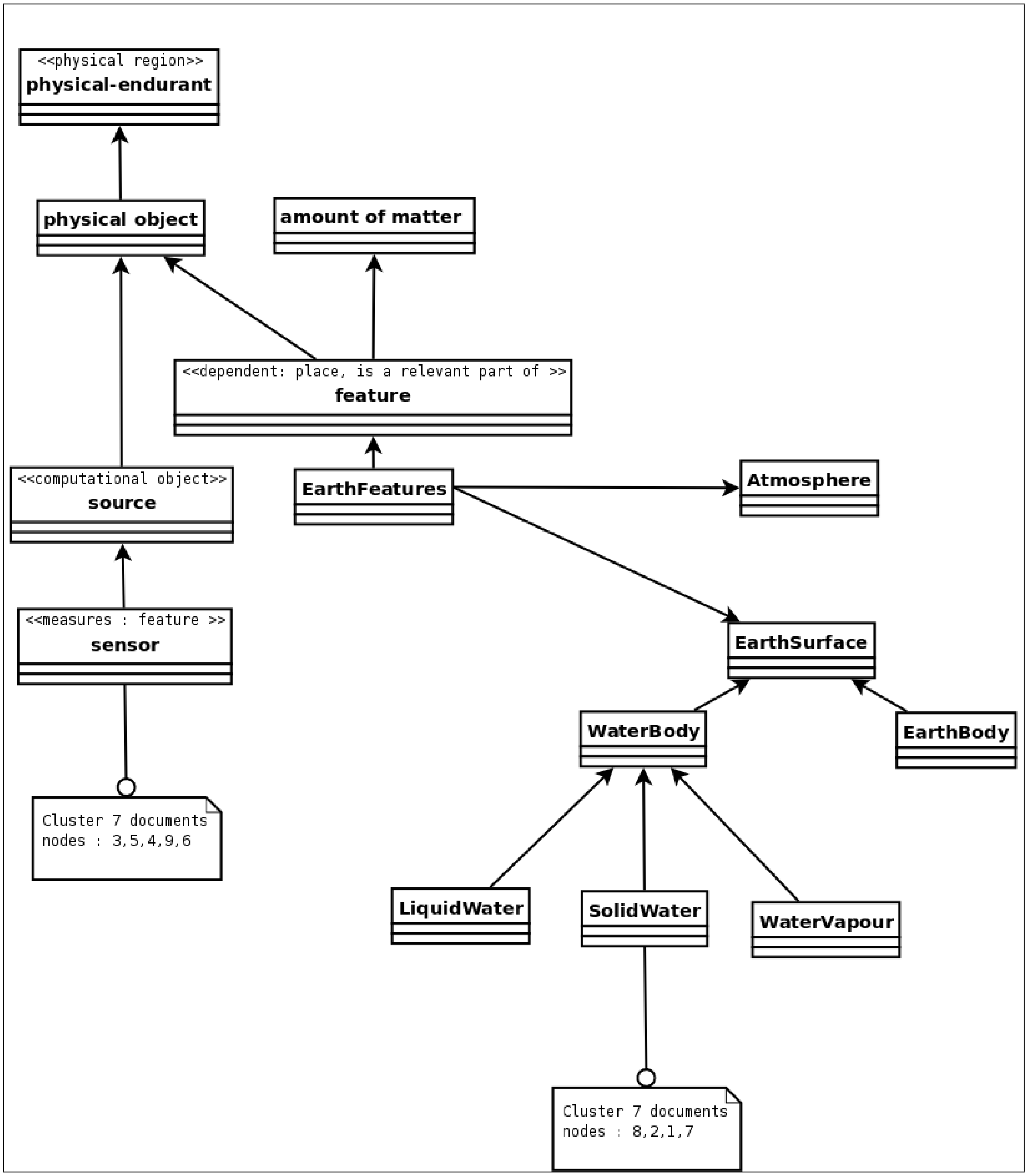


Figure 5.11: Instances created from hierarchical cluster tree

5.5 Discussion

Comparison of LSI and LSI-ontology is performed by analysing the similarity distance measures within a cluster and dissimilarities between clusters. Intra-cluster similarity plots show that LSI-ontology has overall minimum average distance, compared to LSI. This is a result of including the ontology in the corpus where the terms occurring in the ontology alter the term matrix. An additional document containing only the relevant terms that are found in the water ontology create relationships between all water related concepts that LSI on its own is not able to compute. LSI forms a relationship between any two terms if there is a co-occurring term between them. These similarity results indicate that LSI-ontology forms clusters that contain similar documents, so minimum average distances indicate that 'tighter' clusters are formed. The ontology for example forms links between ice, liquid and water vapour from the definition of the water concept. The ontology document therefore adds additional meaning and the largest cluster that was formed using LSI is broken up into different clusters and the documents are re-distributed and more documents are allocated to other clusters. It can be seen from the silhouette plots that LSI-ontology forms two majority clusters and the surrounding clusters [11,12,13,14] are well separated from the rest.

The two majority clusters are used to form a hierarchical cluster tree where they are further broken down into sub-clusters. From cluster 5 it can be seen that ice-related concepts can be easily identified, which can be attributed to the addition of the ontology. Using only LSI the terms 'antarctic' and 'glaciology', cannot easily be associated with ice or snow. The ice-related documents in the respective nodes can be used to instantiate the class 'SolidWater' as is seen in figure 5.11. Figure 5.11 is an illustration of the class hierarchy, showing the relationship between the measured features, in this case EarthSurface related features, and the sensor which measures the feature. The definition and relationships are found in the complete ontology. The second set of documents which are used to create instances, formed by the cluster tree, are sensor related with themes: 'measure' and 'instrument'. The cluster tree shows that by sub-clustering, more theme specific clusters can be formed. The higher the branches formed, the more specific the the theme of the document set. These themes are used for slotting the documents in a particular ontology class, to populate the knowledge base. Storing the documents in the ontology as instances provides a more homogeneous form of storing the documents compared to the original portal arrangement. In addition to this, the ontology also captures relationships between concepts, for example a sensor that measures a feature of a specific type.

5.6 Conclusion

In this chapter LSI is computed in a k-dimension space derived by combining the Everett method and singular value variance computation. Using this reduced latent space, LSI and LSI-ontology are implemented and compared. The performance of LSI and LSI-ontology is compared by evaluating the clusters formed. Evaluation is based on intra-cluster distances and inter-cluster dissimilarity based on Euclidean distance measures. The LSI-ontology results show higher intra-cluster similarity with unrelated clusters separated by larger distances. Finally, using the clusters formed by LSI-ontology specific themes of document groupings are identified. The two sub-clusters which have specific themes that can be readily used to create ontology instances are: sensor documents and ice-related documents. It is shown that the ontology can be used to create meaningful associations between terms, which affect the document clustering, and that the documents can be further classified using cluster trees to store them categorically in the ontology as instances.

CHAPTER VI

CONCLUSION AND RECOMMENDATIONS

In this chapter we present a summary of the research and the accompanying findings. A critical evaluation of the work is presented, which highlight the limitations of the work. Finally the future recommendations and conclusion are presented.

6.1 Summary of research

The research addresses the problem of resource discovery for sensor data published on the web. Information extraction algorithms is used, which forms term and document relationships by decomposition of the term-frequency matrix of the corpus. A comparison is made between the original algorithm, LSI, and the modified algorithm which combines LSI with an ontology to guide the document clustering based on the domain modelling represented by the ontology. The research challenges addressed are:

- Extracting knowledge from published resources using text mining technique

- To investigate the improvement in the information extraction, by introducing a bias towards the resources of interest that contain sensor data by investigation the distances measured within the clusters.

- Evaluate the clustering improvement by introducing the bias, in the form of an ontology representing domain knowledge.

- Automatically creating class instances in the ontology using the clustered resources.

6.2 Summary of findings

Latent Semantic Indexing is used as information extraction algorithm in order to classify sensor resources published on the web. The GCMD portal is used as the test case. This method is combined with an ontology to form the modified LSI-ontology. Latent space reduction of LSI is derived using a combination of singular value variance computation and a previously proposed Everett method. Implementation of both algorithms for the purpose of performance comparison is carried out. The results show that clustering using k-means clustering produces better results for LSI-ontology. This is determined by evaluating the Euclidean distances between clusters, where a high distance measure means high dissimilarity of cluster. The second measure is intra-cluster distances, where low average distance means high similarity of the documents occurring within the cluster. In both cases the LSI-ontology had better overall results. From the clusters formed using LSI-ontology, hierarchichal clustering is performed to further breakdown the themes found in the parent clusters to form sub-clusters. The smaller, more specific groups of clusters are used to populate the ontology as class instances. The two groups of documents that are used for instance creation are sensor related and solid water related.

6.3 Critical Evaluation

It can be seen from the results of the clustering that even though LSI-ontology tends to have lower intra-cluster average distances, the cluster separation distances only improve for certain clusters. The internal criterion improves, which is seen by the re-allocation of documents to the cluster, however the partitioning results improve for certain clusters. It can also be seen that the hierarchical cluster tree can be further broken down into sub-clusters in order to isolate more groups of topic-specific documents that can be used for class instantiation. Another limitation of the research findings is that the evaluation is based on internal criterion, by measuring the cluster distances. An additional useful criterion would be to investigate the global minimum of the k-means objective function. This is useful in the case where there are outliers that occur in the cluster formations. Focusing on the global minimum takes a closer look at the initialisation of the cluster seed. Another limitation is in the implementation phase of the research, where manual switching between operation systems was necessary. The document text processing and term frequency was performed in linux, using c++ language and the open source gcc compiler. The LSI implementation was implemented in Matlab, which uses a windows platform. A solution to this problem could be to use two computers simultaneously and create a workflow, where the results are transferred between operating systems using a script to automate the process.

6.4 Future research and recommendations

The results presented are complete and conclusive, however there is a possibility for further work to be done. The first phase of the process is performed on linux and later on Windows OS for the use of MATLAB. A more automated future implementation would improve the speed of the process and require less user intervention. A possible area of research could be in implementation in a parallel processing environment to ease memory overloading for a large document corpus. The second area is in the clustering phase: Another clustering algorithm can be compared to investigate if there is an improvement in the cluster separation distances obtained by LSI-O versus the LSI. A second portal can also be used to further evaluate the performance of the LSI-ontology. The use of a second portal would mean a second ontology would be used to capture and represent the portal information. The latent space reduction could also be investigated for different document set from a different portal.

The second test collection which can in the future be used as the input for information modelling is the Geoportal, which is derived from the European Space agency (ESA) Earth Observation Community Portal and serves as a single point access to information and services published by the Global Earth Observation System of Systems (GEOSS) GEOSS and access to other geospatial community portals. The difference in structure between the two portals is that the GCMD structures the resources according to measured feature, eg. snow and rainfall and the Geoportal according to area of social benefit, eg. climate, health and disasters. For both portals the resource containing water related information can be considered for classification and direct comparison of the results.

6.5 Conclusion

The objective of the work done is to compare the performance of a proposed approach to the problem of information retrieval. A modification of an existing algorithm, LSI is done by combining with an ontology. Document classification is performed using k-means clustering and the results for both LSI and LSI-ontology are compared. The performance of LSI-ontology shows that documents with higher similarity measures are clustered together, in comparison to LSI. Inter-cluster separations are also slightly wider for LSI-ontology than for LSI. The clusters formed using LSI-ontology are further used to populate the ontology as a knowledge representation of the documents found in the GCMD portal.

APPENDIX A

CONFERENCE PAPER

During the project, selected topics of the work were used to write a conference paper. The paper focus on illustrating the implementation of the modified LSI-ontology algorithm. The focus was on the improvement in document-document clustering and the term groupings that result from including the ontology. The paper was accepted for publication in the 2008 IEEE International Conference on Systems, Man and Cybernetics, October 2008, which was held in Singapore.

The full reference of the publication: Wabo Majavu and Terence T van Zyl and Tshilidzi Marwala. **Classification of Web Resident Sensor Resources using Latent Semantic Indexing and Ontologies**. *2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, October 2008.* ISBN978-1-4244-2384-2.

APPENDIX B

WEB RESIDENT RESOURCE ONTOLOGY VISUALISATION

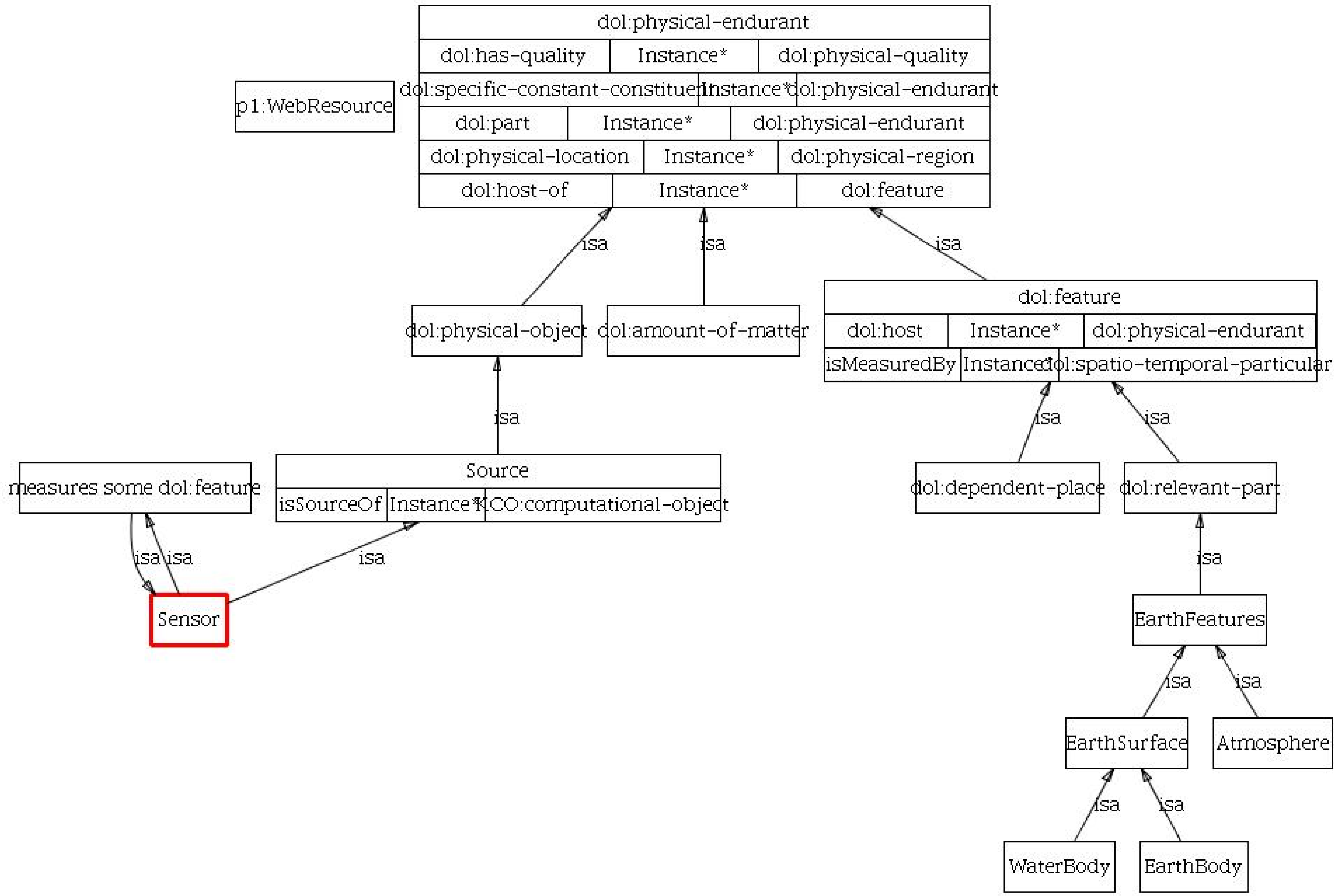


Figure B.1: Basic Relationships between ontological concepts

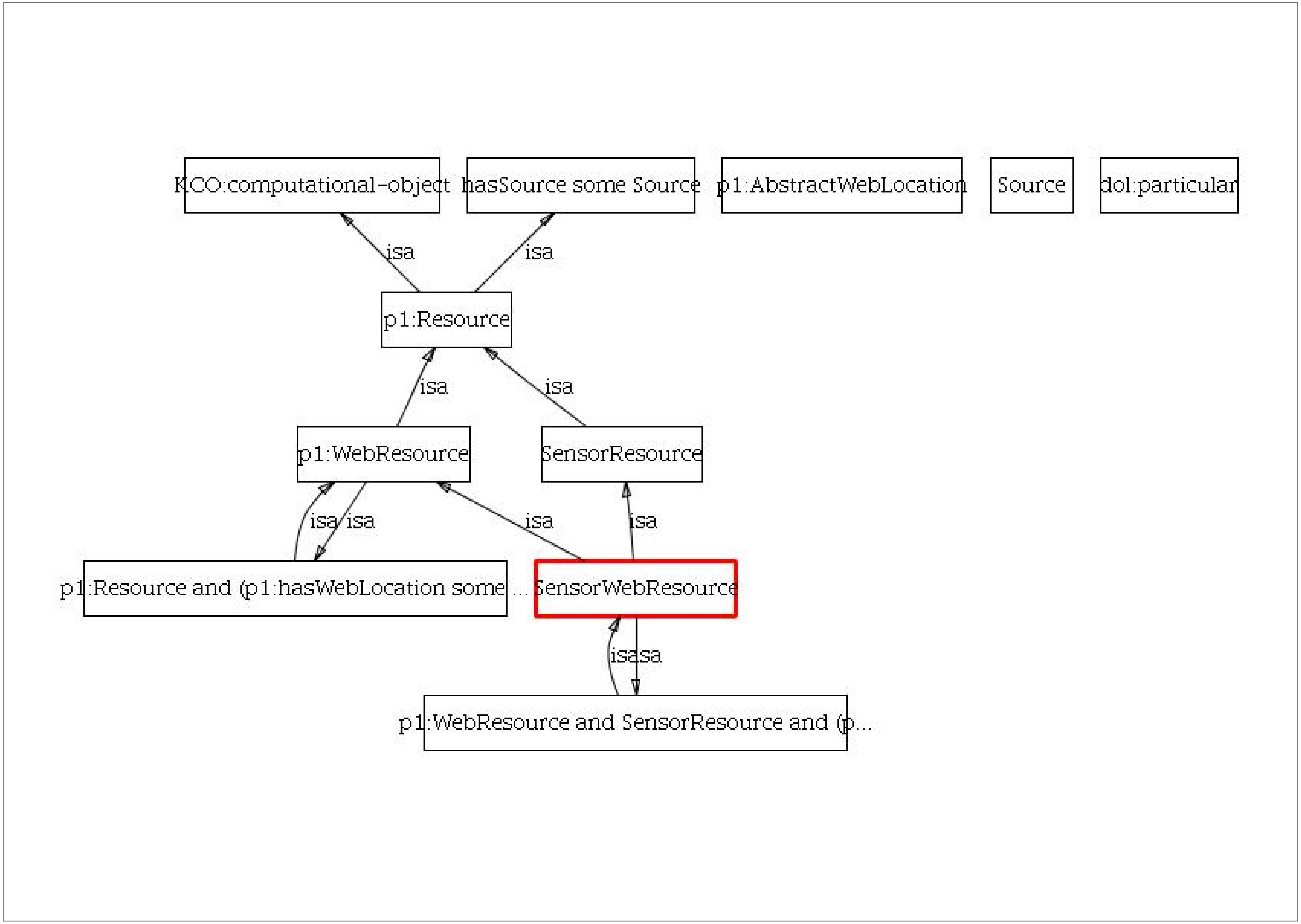


Figure B.2: SensorWebResource class hierarchy

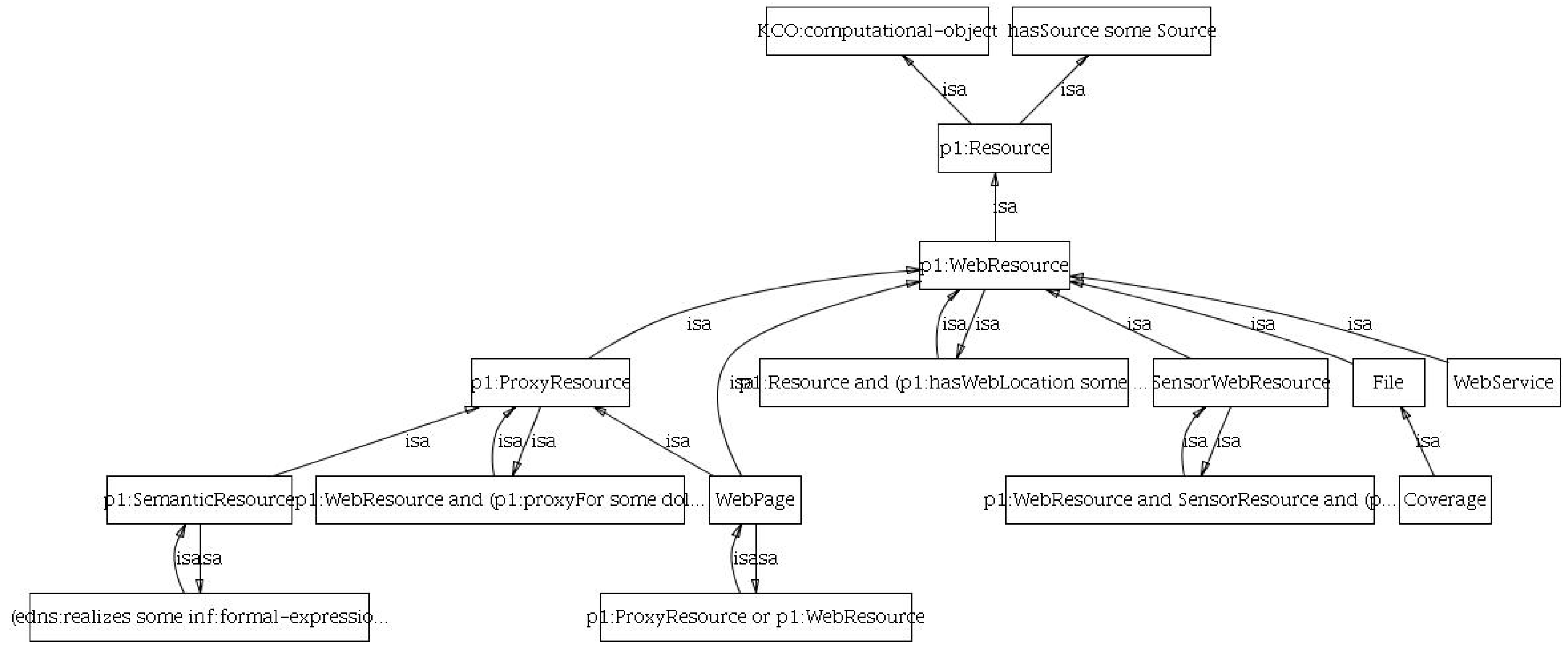


Figure B.3: Extended DOLCE Web Resource descriptions

REFERENCES

[1] Aleman-Meza, B., Halaschek, C., Arpinar, I., and Sheth, A., "Context-aware semantic association ranking," in *Proceedings of SWDB*, pp. 37–50, 2003.

[2] Alesheikh, A. A., Ghorbani, G. M., and Mohammadi, H., "Design and implementation of sensor metadata on internet," in *Remote Sensing*, 2004.

[3] Attardi, G., Gulli, A., and Sebastiani, F., "Automatic web page categorization by link and context analysis," in *Proc. European Symposium on Telematics, Hypermedia and Artificial Intelligence THAI'99)*, pp. 105–119, 1999.

[4] Barfourosh, A. A., Nezhad, H. R., and Perlis, D., "Information retrieval on the world wide web and active logic: A survey and problem definition," Tech. Rep. CS-TR-4291, University of Maryland, 2004.

[5] Beckett, D., "Rdf/xml syntax specification (revised)." In W3C Recommendation, 2004.

[6] Belkin, N. and Croft, W., "Information filtering and information retrieval : two sides of the same coin?," in *Communications of the ACM*, pp. 29–38, 1992.

[7] Berners-Lee, T. and Lassila, J. H. O., "The semantic web," *Scientific American*, no. 284, pp. 34–43, 2001.

[8] Berry, M., Dumais, S., and O'Brien, G., "Using linear algebra for intelligent information retrieval," in *SIAM: Review*, pp. 573–595, 1992.

[9] Botts, M., Robin, A., Davidson, J., and Botts, M., "Opengis sensor web enablement architecture document," Tech. Rep. OGC 06-021, Open Geospatial Consortium, 2006.

[10] Bowman, C. M., Danzing, P., and Manber, U., “Scalable internet resource discovery: Research problems and approaches,” in *Communications of the ACM*, pp. 98–107, 1994.

[11] Chakrabarti, S., den Berg, M. V., and Dom, B., “Focused crawling :a new approach to topic-specific web resource discovery,” in *Proc. Eighth International Conference on The World-Wide Web*, 1999.

[12] Cherry, S., “Some comments on singular value decomposition analysis,” in *Journal of Climate 9: 20032009.*, pp. 2003–2009, 1996.

[13] Cox, S., “Opengis observations and measurements part 1: Observation schema,” Tech. Rep. OGC 07-022r1, Open Geospatial Consortium, 2007.

[14] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Furnas, G. W., and Harshman, R., “Indexing by latent semantic analysis,” *Journal of the American Society For Information Science*, vol. 41, pp. 391–407, 1990.

[15] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Furnas, G. W., and Harshman, R., “Indexing by latent semantic analysis,” in *Journal of the American Society For Information Science*, pp. 391–404, 1990.

[16] Delin, K. A., “The sensor web: A macro-instrument for coordinated sensing,” *Sensors 2002*, vol. 2, pp. 270–285, 2002.

[17] D.Li, “Geospatial sensor web and self-adaptive earth predictive systems,” *NASA AIST PI Conference*, 2007.

[18] Everitt, B. and Dunn, G., “Applied multivariate data analysis,” in *Arnold, 2nd edition.*, 2001.

[19] Frakes, W. and Baeza-Yates, R., "Information retrieval: Data structures and algorithms," in *Prentice-Hall*, 1992.

[20] Frias-Martinez, E., Chen, S. Y., Macredie, R. D., and Xiaohui, L., "The role of human factors in stereotyping behavior and perception of digital library users: a robust clustering approach," in *User Modeling and User-Adapted Interaction*, pp. 305–337, 2007.

[21] Guillaume, D., Futrelle, J., McGrath, R., and Plante, R., "Digital library technology for locating and accessing scientific data," *In ACM Digital Libraries '99*, pp. 188–194, 1999.

[22] Henderson, S., "Genre, task, topic and time: facets of personal digital document management," in *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: making CHI natural*, p. 7582, 2005.

[23] Hitzler, P., Studer, R., and Sure, Y., "Description logic programs: A practical choice for the modelling of ontologies," in *1st Wokshop on Formal Ontologies Meet Industry (FOMI'05)*, 2005.

[24] Jarvelin, K., "An analysis of two approaches iin information retrieval: From frameworks to study designs," in *In First Argentinian symposium on artificial intelligence*, pp. 971–986, 2007.

[25] Jenkins, C., Jackson, M., Burdon, P., and Wallis, J., "Automatic rdf metadata generation for resource discovery," in *Computer Networks*, vol. 31, pp. 1305–1320, 1999.

[26] Jones, C. B., "Spatial information retrieval and geographical ontologies an overview of the spirit project," in *Proceedings of the 25th annual international*

ACM SIGIR conference on Research and development in information retrieval, p. 387388, 2001.

[27] Kao, A. and Poteet, S., "Text mining and natural language processing introduction for the special issue," in *Boeing Phantom Works, SIGKDD Explorations*, vol. 7, 2005.

[28] Khan, M. and Khor, S., "Web document clustering using a hybrid neural network," in *Applied Soft Computing*, vol. 4(4), p. 423432, 2004.

[29] Klassmann, A., Offenga, F., Broeder, D., Skiba, R., and Wittenburg, P., "Comparison of resource discovery methods," in *5th International Conference on Language Resources and Evaluation LREC 2006*, 2006.

[30] Koller, D. and Sahami, M., "Hierarchically classififying documents using very few words," in *Proc. Fourteenth International Conference on Machine Learning*, 1997.

[31] Kontostathis, A. and Pottenger, W. M., "A framework for understanding lsi performance," *Information Processing and Management*, vol. 42(1), pp. 56–73, 2006.

[32] Kossala, R. and Blockeel, H., "Web mining research : A survey," *SIGKDD Explorations*, vol. 2(1), pp. 1–15, 2000.

[33] Lagoze, C., "From static to dynamic surrogates: Resource discovery in the digital age," *D-Lib Magazine*, p. http://www.dlib.org/dlib/june97/06lagoze.html, 1997.

[34] Lausen, H., Stollberg, M., Hernandez, R. L., Ding, Y., Han, S., and D.Fensel, "Semantic web portals state of the art survey," in *Journal of Knowledge Management*, vol. 9(4), pp. 40–49, 2005.

[35] Liang, S., Arie, C., and Tao, V., "A distributed geospatial infrastructure for sensor web," in *Computers and Geosciences*, p. 221231, 2005.

[36] Lin, C. Y., "Assembly of topic extraction modules in summarist," in *In proceedings of the AAAI spring symposium on intelligent text summarization*, pp. 23–25, 1998.

[37] Manning, C., Raghavan, P., and Schutze, H., *Introduction to Information Retrieval.* Cambridge University Press., 2008.

[38] McGuinness, D., "Ontologies come of age," in *In Spinning the Semantic Web Bringing the World Wide Web to its Full Potential*, MIT Press., 2003.

[39] Melnik, S. and Decker, S., "A layered approach to information modeling and interoperability on the web," in *Proc. Workshop on the Semantic Web 4th European Conference on Redsearch and Advanced Technology for Digital Libraries*, 2000.

[40] Meshkova, E., Riihijrvi, J., Petrova, M., and Mhnen, P., "A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks," in *Computer Networks*, vol. 52, pp. 2097–2128, 2008.

[41] Moodley, D. and Simonis, I., "Aa new architecture for the sensor web: the swap framework," in *Semantic Sensor Networks Workshop ISWC'06)*, 2006.

[42] Na, A. and Priest, M., "Opengis sensor observation service implementation specification," Tech. Rep. 0GC 06009r1, Open Geospatial Consortium, 2006.

[43] Nang, J. and Park, J., "An efficient indexing structure for content based multimedia retrieval with relevance feedback," in *Proceedings of the 2007 ACM symposium on Applied computing*, p. 517534, 2005.

[44] Noy, N. F. and McGuinness, D., “Ontology development 101: A guide to creating your first ontology,” Tech. Rep. KSL-01-05, Stanford KSL Technical Report, 2000.

[45] Pant, G., Srinivasan, P., and Menczer, F., “Crawling the web,” in *Web Dynamics in New York: Springer-Verlag*, vol. In M. Levene and A. Poulovassilis Editors, p. 153178, 2003.

[46] Pretschner, A., “Ontology based personalized search,” in *Web Intelligence and Agent System*, pp. 219–234, 2003.

[47] Rajaraman, K. and Tann, A., “Topic detection, tracking and trend analysis using self-organizing neural networks,” in *In Proceedings of the PAKDD*, p. 102107, 2001.

[48] Salton, G., Wong, A., and Yang, C., “A vector space model for automatic indexing,” in *ACM Communications*, vol. 18(11), pp. 613–620, 1975.

[49] S.Chakrabarti, “Recent results in automatic web resource discovery,” in *wabo*, p. wabo, wabo.

[50] Schwartz, M. F., Emtage, A., Kahle, B., and Neuman, B. C., “A comparison of internet resource discovery approaches,” in *Computing Systems*, vol. 5(4), 1992.

[51] Sebastiani, F., “A tutorial on automated text categorisation,” in *In First Argentinian symposium on artificial intelligence*, pp. 7–35, 199.

[52] Shen, Y., Lee, D. L., and Zhang, L. W., “A distributed search system based on markov decision processes,” in *In Proceedings of the ICSC 99 conference*, 1999.

[53] Siring, E., Hendler, J., and Parsia, B., "Semi-automatic composition of web services using semantic descriptions," in *In WSMAI2003*, 2003.

[54] Tan, A. H., "Text mining: The state of the art and the challenges," in *In Proc. of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, p. 6570, 1999.

[55] Toms, E., "Understanding and facilitating the browsing of electronic text," in *International Journal of Human-Computer Studies*, vol. 52(3), pp. 423–452, 2000.

[56] Wang, M. and Nie, J., "A latent semantic structure model for text classification," in *ACM-SIGIR Workshop on mathematic/formal methods in information retrieval*, 2003.

[57] Yu, B., Xu, Z., and Li, C., "Latent semantic analysis for text categorization using neural network," in *Knowledge-Based Systems*, 2008.

[58] Zhao, P., Chen, A., Liu, Y., Di, L., Yang, W., and Li, P., "Grid metadata catalog service-based gc web registry service," in *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, 2004.

[59] Zhao, P. and Di, L., "Semantic web service based geospatial knowledge discovery," in *Geoscience and Remote Sensing Symposium IGARSS'06*, 2006.