The Islamic University – Gaza
Research & Graduate affairs
Faculty of Information Technology
Master of Information Technology

Investigating Approaches to Enhance Document Clustering by exploiting Background Knowledge in WordNet and Wikipedia

Submitted by:
Rami Hassouna Nafee

Supervised by:
Dr. Iyad M. Alagha

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology

2015

بسم الله الرحمن الرحيم

"وقل ربي زدني علما"

صدق الله العظيم

Dedication

To my beloved father

To my beloved mother

To my beloved wife and my sons

To my beloved aunts

To sister and brothers

To my best friends

Acknowledgment

Praise and thanks is to Allah always and never

I owe deep debt of gratitude to my supervisor, DR. Iyad Alagha , for his guidance, enthusiasm and continuous follow-up and continuous encouragement of my work through my thesis. This thesis would not have been possible without his support to me.

I would like to express my sincere thanks to my teachers of the Faculty of Information Technology who helped me during different courses.

Last but not least, Thanks of a special kind to my family: my parents, my wife, my aunts, my brothers, and my sister for their support me all along my education.

Abstract

Clustering is one of the main data analysis techniques. Document clustering generates clusters from the whole document collection automatically and it is used in numerous applications, including market research, pattern recognition, data analysis, and image processing. Traditional techniques of document clustering do not consider the semantic relationships between words when assigning documents to clusters. For instance, if two documents talking about the same topic but by using different words (which may be synonyms or semantically associated), these techniques may assign documents to different clusters. Previous research has approached this problem by enriching the document representation with the background knowledge from an ontology or a controlled vocabulary such as Wordnet. This research builds on previous efforts and provides a thorough investigation on the use of controlled vocabularies such as WordNet and knowledge resources such as Wikipedia to enhance document clustering. The contribution of this research is twofold:

First, it provides a thorough investigation on the value of using WordNet to enhance document clustering: previous researches which explored the use of WordNet for document clustering often showed conflicting results: some efforts claim that WordNet has the potential to improve the performance of the clustering by helping to identify synonyms and semantically related words in the document collection. Other researches claim that WordNet provides little or no enhancement on the clustering results. In this research, we will try to experimentally resolve this conflict between the two teams, and explain why WordNet could be useful in some cases while not in others, and what factors can influence the value

of the WordNet. We have conducted several experiments in which we tested the use of WordNet for document clustering over different testing conditions such as different data sets, different similarity measures and different settings for the clustering algorithm. Results have shown that different experimental settings will result in different results, and that the influence of WordNet on the clustering results varies based on the used settings. The importance of these results is that they can inform the designers of experiments, who are willing to use WordNet for document clustering, of the best settings they should use in order to obtain the ultimate benefit from WordNet, For instance, using the Reuters dataset, the clustering with synonyms gave the best results (F-score =0.77 and purity =0.64), followed by the clustering with similarity scores (F-score=0.70, Purity=0.59), followed by the clustering without any semantics (F-score=0.64, Purity=0.57).

Second, this thesis presents a novel approach to enhance document clustering by exploiting the semantic knowledge contained in Wikipedia. It uses the link structure of Wikipedia to measure the semantic relatedness between terms and use the similarity scores to enhance the document's representation vector. The proposed approach differs from related efforts which also used Wikipedia for document clustering in two aspects: first, it uses a similarity measure that is modelled after the Normalized Google Distance which is a well-known and low-cost method of measuring term similarity. Second, it is more time efficient as it applies an algorithm for phrase extraction from documents prior to mapping terms to Wikipedia. Our approach was evaluated by being compared with different methods from the state of the art using two

different datasets. Empirical results showed that our approach improved the clustering results as compared to other similar approaches, According to the F-score measure, for the Reuters dataset, our method (Wikipedia) and Hotho et al's method (WordNet) achieve 31% and 9% respectively, for the OHSUMed dataset, our method and Hotho et al's method achieve 27% and 4% respectively.

Keywords: Document Clustering, WordNet, Wikipedia, Semantic Similarity Measures, Synonyms, k-means Algorithm , Vector Space Model, Apriori Algorithm, Frequent Item Sets, Normalized Google Distance.

عنوان البحث

التحقيق في طرق لتحسين تصنيف الملفات عن طريق استغلال المعرفة الخلفية في وردنت و ويكيبيديا

ملخص البحث

التصنيف هي واحدة من تقنيات تحليل البيانات الرئيسية وتصنيف المستندات يولد مجموعات من مجموعة كبيرة من المستندات ويتم استخدامه في تطبيقات عديدة بما في ذلك أبحاث السوق، والتعرف على الأنماط، وتحليل البيانات، ومعالجة الصور.

التقنيات التقليدية لتصنيف المستندات في مجموعات بحيث تكون كل مجموعة تحتوي على المستندات المتشابهة لا تستغل العلاقات الدلالية بين الكلمات في المستندات. على سبيل المثال، إذا وثيقتين تتحدث عن نفس الموضوع ولكن باستخدام كلمات مختلفة (ربما تكون مترادفات أو مترابطة في المعنى) هذه التقنيات تقوم بوضع هذه المستندات في مجموعات مختلفة.

الأبحاث السابقة تعاملت مع هذه المشكلة عن طريق استخدام المعرفة الخلفية من الانطولوجيا أو قاموس مثل وردنت، ويستند هذا البحث على الجهود السابقة بإجراء تحقق شامل في استخدام قاعدة بيانات معجمية مثل وردنت والموسوعة المعرفية مثل ويكيبيديا لتحسين تصنيف المستندات ومساهمة هذا البحث في شقين :

أولا، إجراء تحقيق شامل حول قيمة استخدام وردنت لتحسين تصنيف المستندات: البحوث السابقة التي استخدمت وردنت لتصنيف المستندات غالبا ما أظهرت نتائج متضاربة، بعض الجهود وجدت أن وردنت لديه القدرة على تحسين أداء التصنيف عن طريق المساعدة على تحديد المترادفات والكلمات ذات الصلة لغويا وأبحاث أخرى وجدت أن وردنت يوفر تحسين ضئيل أو معدوم على نتائج تصنيف المستندات في مجموعات.

في هذا البحث، سوف نحاول بتجربة لحل هذا الاختلاف بين الفريقين، وتوضيح لماذا الوردنت هي مفيدة في بعض الحالات وغير مفيدة في حالات اخرى وما هي العوامل التي تؤثر على قيمة وردنت.

لقد أجريت العديد من التجارب باستخدام وردنت لتصنيف المستندات في مجموعات بظروف مختلفة مثل مجموعات مختلفة من المستندات واستخدام مقاييس التشابه المختلفة واعدادات مختلفة وقد أظهرت النتائج أن إعدادات التجربة المختلفة سوف تؤدي إلى نتائج مختلفة، وأن تأثير وردنت على النتائج تختلف تبعا للإعدادات المستخدمة. أهمية هذه النتائج هي أنها يمكن أن تعلم من هم على استعداد لاستخدام وردنت في تصنيف المستندات الى مجموعات لأفضل الإعدادات التي يجب استخدامها من أجل الحصول على الاستفادة القصوى من وردنت.

على سبيل المثال، وذلك باستخدام مجموعة بيانات رويترز، أعطى التصنيف مع المترادفات أفضل النتائج (F-score =0.77 and purity =0.64)، يليه تصنيف مع مقاييس التشابه (F-score=0.70, Purity=0.59)، يليه تصنيف دون أي دلالات (F-score=0.64, Purity=0.57).

ثانيا، تقدم هذه الأطروحة نهجا جديدا لتحسين تصنيف المستندات المتشابهة من خلال استغلال المعرفة الدلالية الواردة في ويكيبيديا. ويستخدم بنية الارتباط من ويكيبيديا لقياس الصلة الدلالية بين المصطلحات واستخدام عشرات التشابه لتحسين تمثيل المستندات.

النهج المقترح يختلف عن الجهود السابقة ذات الصلة والتي تستخدم أيضا ويكيبيديا في تصنيف المستندات في جانبين هما:
أولا، فإنه يستخدم مقياس التشابه Normalized Google Distance وهو معروف ومنخفض التكلفة لقياس

التشابه بين الكلمات.
ثانيا، لاستغلال الوقت بفعالية سنطبق خوارزمية لاستخراج العبارة من المستند قبل استخدام الويكيبيديا.

تم تقييم طريقتنا باستخدام مجموعتين من المستندات تم استخدامهم في ابحاث سابقة وأظهرت النتائج التجريبية أن نهجنا حسنت نتائج التصنيف بالمقارنة مع الطرق الأخرى المماثلة.

وفقا لمقياس F-score، لمجموعة البيانات رويترز، لدينا وسيلة (ويكيبيديا) وطريقة Hotho (وردنت) تحقق ٣١٪ و ٩٪ على التوالي، لمجموعة البيانات OHSUMed، أسلوبنا وطريقة Hotho لتحقق ٢٧٪ و ٤٪ على التوالي.

كلمات مفتاحية: تصنيف المستندات، وردنت، ويكيبيديا، مقاييس التشابه الدلالي.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

VSM	Vector Space Model
TF-IDF	Term Frequency - Inverse Document Frequency Model
POS	part-of-speech tags
BOW	Bag of Words
NGD	Normalized Google Distance Measure

Chapter 1

Introduction

1.1 Overview

With the increase of information resources such as publications, books, web pages in various domains, there is a need to arrange these resources in an easy manner by organizing them in groups in a process called clustering. Clustering groups a collection of objects into meaningful sub-groups, where each group represents a similar objects[1, 2]. Document clustering generates clusters from the whole document collection automatically and it is widely used in a variety of applications including pattern recognition, data analysis, marketing, economics and image processing [3, 4]. An example of a clustering process is depicted in figure 1.1.

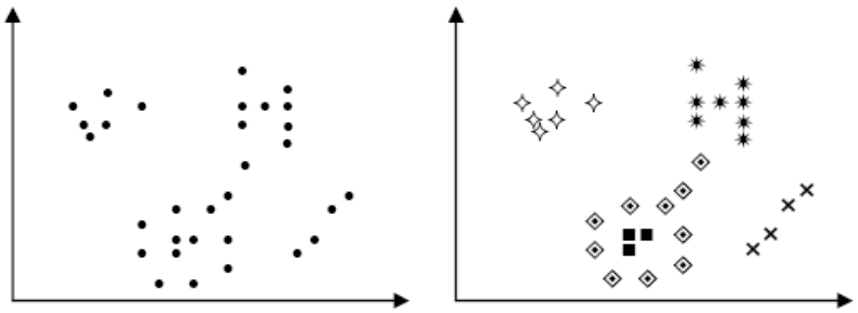

Figure 1.1 : A: collection of Documents, B: Clustering of documents

Existing research on document clustering presented a large number of clustering techniques. Most of these techniques deal with a document as a bag of words and often rely on the existence of keywords and the number of occurrences to cluster documents.

These techniques, however, do not take into account the fact that the keywords may have some semantic proximity between each other depending on the context[5]. The word “semantics” is related to the word syntax. In most languages, syntax is how you say something, where semantics is the meaning behind what you have said[6], which shows of the relation between context and meaning. For example, the words “camel” and “desert” are semantically related as we know that the camel

lives in the desert. However, a clustering technique relying on keyword-based matching will ignore relationships between the terms that do not occur literally.

There is an emergent need to increase the quality of clustering by integrating semantics rather than syntactic characteristics of text in order to benefit of the relation between context and meaning. Many efforts proposed to enhance the document's representation by measuring the semantic similarity between terms. For this purpose, domain ontologies can be used as a background knowledge. A domain ontology defines the terms used in a particular domain of knowledge and the relationships between these terms.

There exist plenty of research efforts which used ontologies to measure semantic similarities between concepts for various purposes such as sense disambiguation [7], information extraction and retrieval[8, 9], and to enhance document clustering. For example, some studies[10, 11] used the WordNet or other dictionaries to determine word synonyms and other types of semantic relations such as hyponymy, hyponymy and antonymy. This information is important to precisely measure the semantic similarity between terms and thus enhance the clustering results by assigning documents that have related keywords to the same clusters.

The first part of this research aims to explore the use of controlled vocabularies such as WordNet to enhance document clustering. An experimental study will be conducted to investigate the factors that affect the use of WordNet for measuring the similarity between the document terms.

In the second part of this research, we will propose an approach to enhance document clustering by exploiting the semantic information

obtained from Wikipedia. The similarity between terms in the document collection will be estimated by using a measure that is based on the link structure of Wikipedia. Subsequently, the document's representation will be augmented with the similarity scores obtained from Wikipedia.

1.2 Research Contributions

This research contributes to a better understanding of the field of semantically-enhanced document clustering by the following:

1. **Resolve the conflict over the influence of using WordNet for enhancing the document clustering**: First, it investigates the influence of using WordNet as a background knowledge for document clustering. Existing approaches that used WordNet for document clustering often report conflicting results: while some researches show that WordNet has the potential to improve the clustering results by enhancing the document's representation [12-14], other approaches claim that WordNet results in little or no improvement, or may even degrade the clustering performance [15-17]. To resolve this conflict, we conducted an experimental study in which WordNet was used for document clustering across different testing conditions and experimental settings.

The main objective is to evaluate WordNet as a background knowledge in improving document clustering process by using semantic similarity measures and synonyms between terms instead of the use of traditional method.

The hypothesis are the results of the WordNet depends on two factors the dataset and similarity measures, whereas the most previous studies which we have mentioned used the same dataset and there was differences in

the results between the utility and non-utility in improving clustering process efficiency. This study is fully reported in Chapter 3.

2. **Propose a new approach to enhance document clustering by exploiting the semantic knowledge contained in Wikipedia**. Our approach differs from related efforts in two aspects: first, unlink others who built their own methods of measuring similarity through the Wikipedia categories; our approach uses a similarity measure that is modeled after the Normalized Google Distance which is a well-known and low-cost method of measuring term similarity. Second, it is more time efficient as it applies an algorithm for phrase extraction from documents prior to matching terms with Wikipedia. The proposed approach is discussed in detail in Chapter 4.

1.3 Statement of Problem

The research problem explored in this thesis is twofold:

First, many efforts proposed the use of controlled vocabularies such as WordNet or domain ontologies to enhance document clustering by measuring the semantic proximity between document terms. These efforts often showed contradictory results, indicating that the existing efforts have failed to reach a conclusive result as to whether these controlled vocabularies can improve document clustering or not, and to explain why these vocabularies can be useful in some cases while not in others.

It is clear from the above discussion that there is a conflict regarding the value of WordNet as background knowledge for document clustering: While some efforts reported that WordNet has the potential to improve the clustering results, others reported that WordNet has little or no impact, or even can introduce noise that hinder the clustering process.

Second, some efforts proposed to use Wikipedia as a background knowledge to incorporate semantics into document clustering. This is motivated by the fact that Wikipedia has a much better coverage than domain-specific ontologies or WordNet. However, existing approaches presented many challenges such as the need to pre-process the whole Wikipedia content prior to the clustering process or the use of application-specific similarity measures whose accuracy is well assessed.

1.4 Objectives

- One objective of this research is to try to resolve this conflict by seeking answers to the following major questions:

  - What potential factors could make WordNet useful in particular situations and while not in others situations?
  - Do the different experimental settings have impact on the clustering performance?
  - How the obtained result can inform the design of WordNet based clustering techniques?

To answer the above questions, we will explore the use of WordNet for document clustering across different experimental conditions. These conditions involve the use of different datasets, different similarity measures and different preprocessing steps. We aim to explore how different combinations of these settings could result in different clustering results. In light of previous researches, we will also try to explain, why WordNet was effective in some case while not in others.

- Second objective of this research is to propose a novel approach to improve document clustering by explicitly incorporating the semantic similarity between Wikipedia concepts into the document's vector space model.

1.5 Importance of research

Document clustering is an essential process for an enormous number of computer applications in different disciplines. Most existing research on document clustering has considered techniques such as keyword concurrences. There is an emergent need to increase the quality of clustering by integrating semantics rather than syntactic characteristics of text. In an endeavor to improve clustering results, this research leverages the recent advances in Semantic Web to determine potential links between document terms. Results of our research contribute to a better understanding of the value of semantics for enhancing documents clustering. We also think that the provided experimental results help to resolve the conflicting results in previous studies regarding the value of controlled vocabularies, such as WordNet, in the clustering process.

1.6 Scope and limitations of the project

1- For the experimental study we conducted, k-means was chosen for document clustering, which is one of the oldest and most widely used clustering algorithm. We used k-means because it is easy to implement and is widely used. We implemented different testing conditions when using WordNet but with the same clustering algorithm which is k-means. Our intention was to test different conditions (e.g. different data sets, different similarity measures, different experimental settings) while unifying the clustering method. Most importantly, we used k-means to make our work comparable with similar works [18] [19] [14] which also used k-means as a clustering method.

2- When measuring the semantic similarity between document terms, we used a subset of similarity measures which are very common (such as

Lin's measure [20], Wu & Palmer's measure [21]) and excluded those that are less common (such as Tversky's measure [22]).

3- In part of our experiment, we used a dataset that we built. The motivation behind defining our own dataset was to assess the clustering approach when using a dataset that covers a specific domain of knowledge rather than using general datasets that have wider coverage.

1.7 Methodology

Methodology of this research is shown in figure 3.1 that comprises of the following steps:

Test1, Test2, Test3 on WordNet, while Test4 on Wikipedia.

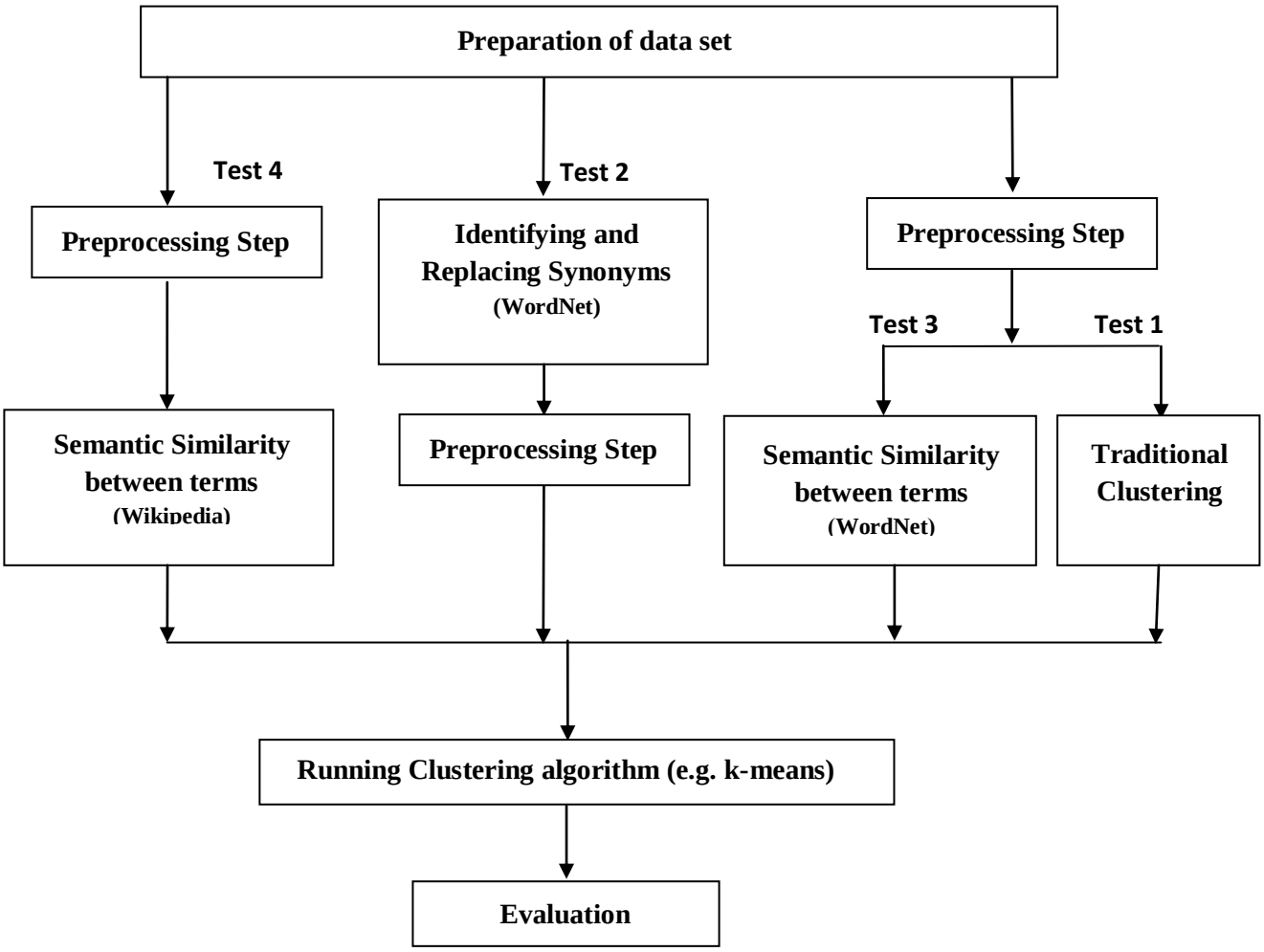

Figure 1.2 : Steps of Methodology

1.8 Thesis Structure

The rest of research is organized as follows: chapter 2 is literature review; chapter 3 is about related works; chapter 4 presents experiment and results about investigating the influence of WordNet on document clustering; chapter 5 presents an efficient approach for semantically-enhanced document clustering by using wikipedia link Structure and chapter 6 is the conclusion and future work.

Chapter 2

Literature Review

Introduction

Clustering is one of the main data analysis techniques and a crucial area by researchers in many fields including data mining, marketing. Importance of document clustering is now widely for better organization and efficient querying of large collection of documents[3]. Document Clustering aims to group among documents in such a way that documents with in a cluster are similar to one another and are dissimilar to documents in other clusters[24]. Traditionally, document clustering approaches mainly uses words, phrases, and sequences from the documents to achieve cluster, but these approaches perform clustering independent of the context[1] [25] [26]. Instead of them, there are approaches integrate domain ontology as background knowledge in document clustering process to exploit the semantics between terms[10] [30].

This chapter aims to review the points of knowledge and concepts that were used by thesis experiments. The chapter is divided into four sections, in section 2.1 we will give definitions about clustering and document clustering concepts and used techniques for clustering, in section 2.2 we will present overview about Ontology and Semantic Web, in section 2.3 we will present Ontology based document clustering, in section 2.4 we will present overview about clustering algorithm, in section 2.5 we will present overview about extraction of frequent phrases algorithm, finally in section 2.4 we will give some conclusions about this chapter.

2.1 Document clustering

Clustering is one of the main data analysis techniques and deals with organizing a set of objects in a multidimensional space into cohesive groups, called clusters for better management and navigation[5].

Clustering is an example of unsupervised learning ,classification refers to a procedure that assigns data objects to a set of classes ,unsupervised means that clustering does not depends on predefined classes and training examples through classifying data objects[3, 23]. Document clustering is useful for many information retrieval tasks such as document browsing, organization and viewing of retrieval results[18].

Many clustering algorithms exist in the literature but difficult to provide a categorization of clustering methods because these categories may overlap, so that a method may have features from several categories, however, the major clustering methods can be classified into the following main categories hierarchical methods, partitioning methods[24].

The partitioning method attempts a flat partitioning of a collection of documents into a predefined number of disjoint clusters[5]. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from group to another, partitioning methods include k-means and k-medoids[24].

Hierarchical methods produce a sequence of nested partitions[5]. The method can be classified as being either agglomerative (bottom-up) or divisive (top-down)[24].

Most of techniques used in document clustering deal with a document as a bag of words without considering the semantics of each document. Traditional algorithms mainly uses features like: words, phrases, and sequences from the documents based on counting and frequency of the features to perform clustering independent of the context[1] [25] [26] [27] [28].They ignore the semantics among words in documents.

2.2 Ontology

Current researches efforts in document clustering started to focus on the development of a more efficient clustering with considering the semantics between terms in documents to enhance the clustering results.

To address specific domain terminologies, Ontology can be used to model the various semantic relations that exist between concepts. An ontology formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts[29]. Relations defined within ontologies represent ways in which classes and individuals can be related to one another.

The semantic similarity have been tested on WordNet and ontology to determine relatedness between terms[10] [30] [11] [31]. Such that, two concepts may belong to two different nodes in an ontology and the distance between their nodes determines the similarity of these two concepts[31].

It has been widely used in information retrieval, sense disambiguation, text segmentation, question answering, recommender system, information extraction and so on. In the next section, we explore the existing semantic similarity measures that use ontology as primary information source.

2.3 Ontology based document clustering

A number of research efforts explored on how to use of dictionary based techniques (e.g. WordNet or domain ontologies) to enhance document clustering by measuring the semantic proximity between document terms [19] [15] [13].

The clustering process performs in three steps, document preprocessing to remove unwanted terms and symbols, document representation to transform each document into a vector of term weights by calculating weights using semantic similarity based on the ontology and clustering of documents.

2.3.1 Ontology-based similarity measures

Several measures have been proposed for determining semantic similarity between terms. These measures include Path-based Counting measures, Information Content measures and Feature-Based measures.

Similarity Measures are :

1- **Path-based Measures :** Measure the similarity between two concepts (i.e., C_1 and C_2) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy [10, 32] . This kind of measures is called as *Edge-based* , that measures contain The Shortest Path based Measure, Wu & Palmer's Measure, Li's Measures, Leakcock& Chodorow's Measure, Mao and Chu's Measure.

2- **Information Content-based Measures**: Measure the more common information between two concepts terms (i.e., C_1 and C_2) that depended on the information content that subsumes them in the ontology, where each concept includes much information[10]. This kind of measures is called as Node-based. That contain Resnik's Measure, Lin's Measure, Jiang's Measure

3- **Feature-based Measure**: Measure the similarity between two terms as a function of their properties or based on their relationships to other

similar terms in the taxonomy. Common features tend to increase the similarity and (conversely) non-common features tend to diminish the similarity of two concepts[32]. That contains Basic Feature, Tversky's measure, Knappe

(the above measures referred to A survey for Semantic Similarity Measures[10] [32] [33] [19]) .

Overview on Similarity Measures:

In our experiment, we used eight wordnet-based measures, is shown below in Table 2.1.

Type	Reference	Module
Path-based Measures	Leacock and Chodorow (1998)	WS4J.runLCH
Path-based Measures	counting nodes in WordNet 'is-a' hierarchies	WS4J.runPATH
Path-based Measures	Wu & Palmer (1994)	WS4J.runWUP
Information Content-based	Jiang and Conrath (1997)	WS4J.runJCN
Information Content-based	Lin (1998)	WS4J.runLIN
Information Content-based	Resnik (1995).	WS4J.runRES
Relatedness measures in WordNet	Banerjee and Pedersen (2002)	WS4J.runLESK
Relatedness measures in WordNet	Hirst and St-Onge (1998).	WS4J.runHSO

Table 2.1 : WordNet-based measures

LCH : This module computes the semantic relatedness of word senses. This method counts up the number of edges between the senses in the 'is-a' hierarchy of WordNet. The value is then scaled by the maximum depth of the WordNet 'is-a' hierarchy. A relatedness value is obtained by taking the negative log of this scaled value[34].

PATH : This module computes the semantic relatedness of word senses by counting the number of nodes along the shortest path between the senses in the 'is-a' hierarchies of WordNet. The path lengths include the end nodes.

WUP : RES module revises the WUP module of measuring semantic relatedness. RES uses an edge distance method by taking into account the most specific node subsuming the two concepts. Here we have implemented the original WUP modul, which uses node-counting.

JCN : This module computes the semantic relatedness of word senses. This measure is based on a combination of using edge counts in the WordNet 'is-a' hierarchy and using the information content values of the WordNet concepts. Their measure, however, computes values that indicate the semantic distance between words (as opposed to their semantic relatedness). In this implementation of the measure we invert the value so as to obtain a measure of semantic relatedness. Other issues that arise due to this inversion (such as handling of zero values in the denominator) have been taken care of as special cases.

LIN : This module describes a method to compute the semantic relatedness of word senses using the information content of the concepts in WordNet and the 'Similarity Theorem'. This module implements this measure of semantic relatedness of concepts.

RES : This module uses the information content of concepts, computed from their frequency of occurrence in a large corpus, to determine the semantic relatedness of word senses. This module implements this measure of semantic relatedness.

LESK : This module proposed that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. LESK extended this notion to use WordNet as the dictionary for the word definitions. This notion was further extended to use the rich network of relationships between concepts present is WordNet. This adapted lesk measure has been implemented in this module.

HSO : This module computes the semantic relatedness of word senses according. In their paper they describe a method to identify 'lexical chains' in text. They measure the semantic relatedness of words in text to identify the links of the lexical chains.

2.3.2 Examples of Ontologies used to enhance document clustering

2.3.2.1 WordNet

Overview

WordNet is an example of ontologies that is widely used as a background knowledge for document clustering. WordNet is the product of a research project at Princeton University [35]. It is a large lexical database of English. In WordNet Nouns, verbs, adverbs and adjectives are organized by a variety of semantic relations into synonym sets (synsets), which represent one concept. Examples of semantic relations used by WordNet are synonymy, autonomy, hyponymy, member, similar, domain and cause and so on. Some relations are used for word form relation and others for semantic relations. These relations will be associated with words and words to form a hierarchy structure, which makes it a useful tool for computational linguistics and natural language processing. It is commonly argued that language semantics are mostly captured by nouns or noun phrases so that most of the researches focus on noun in semantic similarity calculating. There are four commonly used semantic relations

for nouns, which are hyponym/hypernym (is-a), part meronym/part holonym (part-of), member meronym/member holonym (member-of) and substance meronym/substance holonym (substance-of). For example, apple is a fruit (is-a) and keyboard is part of computer (part-of). Hyponym/hypernym (is-a) is the most common relation, which accounts for close to 80% of the relations[10].

Various researches have concentrated on comparing the effects semantic similarity measures of term on document clustering based on Wordnet as ontology.
Recupero and Diego Reforgiato [12] uses Wordnet to perform dimensionality reduction prior to clustering.

Hung et al. [36] uses a hybrid neural network model guided by Wordnet to cluster documents.
Many researches that used WordNet will explain in chapter 3.

2.3.2.2 Domain specific ontologies

There are studies used domain specific ontologies such as MeSH ontology. MeSH ontology defines a taxonomic structure of medical vocabularies. Thus, the similarity measures in these studies were restricted to taxonomic relationships.

Some efforts [19, 33] evaluated the effects of the similarity measures that include four path based similarity measure, three information content based similarity measure, and two feature based similarity measure on PubMed document sets. The result of the evaluation process showed that there is no a certain type of similarity measures that significantly outperforms the others, several similarity measures have rather more stable performance than the others.

Zhu et al. [37] proposed a strategy for clustering the MEDLINE documents based on the semantic information which is derived from the MeSH thesaurus by mapping the document vectors on it. Spectral clustering is used for grouping the documents that based on the integrated similarity matrix. The similarity matrix is used to record both the semantic and content similarities between the documents. Experiment used various 100 datasets of MEDLINE records. The results of Experiment show that integrating the semantic and content similarities outperforms the case of using only one of the two similarities, being statistically significant.

2.4 clustering algorithm

We use k-means as clustering algorithm on the collected datasets in our experiments, k-means is a popular baseline method used by previous researchers on ontology-based clustering algorithms, The algorithm is shown in figure 2.1.

Overview on k-means

We chose the k-means as an example of clustering methods, which is one of the oldest and most widely used clustering algorithm for clustering process to find coherent groups of data, Document clustering employs K-means clustering since its complexity is linear in n, the number of elements to be clustered. K-means is a family of partitional clustering algorithms[18], we programmed K-means in java, to integrate it with any java API.

Before being able to run k-means on a set of text documents, the documents have to be represented as mutually comparable vectors. To achieve this task, the documents can be represented using the tf-idf score. The tf-idf, or term frequency-inverse document frequency, is the most

common weighting method used to describe documents in the Vector Space Model.

After that we are equipped with a numerical model to compare our data where each document as a vector of terms using a global ordering of each unique term found throughout all of the documents. After we have our data model, we have to compute distances between documents. Visual k-means representations, the data consists of plotted points usually use what looks like Euclidian distance; however, in our representation, instead we can calculate the cosine similarity between the two "arrows" of each document vector. Cosine similarity of two vectors is computed by dividing the dot product of the two vectors by the product of their magnitudes.

k-means clustering works by assigning data points to a cluster centroid, and then moving those cluster centroids to better fit the clusters themselves[38].

Basic K-means Algorithm:
1- Select K points as the initial centroids.
2- Assign all points to the closest centroid.
3- Recompute the centroid of each cluster.
4- Repeat steps 2 and 3 until the centroids do not change.

Figure 2.1 : Basic K-means Algorithm

Running an iteration of k-means on our dataset:

We first randomly initialize k number of points to serve as cluster centroids. A common method, employed in my implementation, is to pick k data points and fix the centroid in the same place as those points. Then we assign each data point to its nearest cluster centroid. Finally, we update the cluster centroid to be the mean value of the cluster. The assignment and updating step is repeated, minimizing fitting error until

the algorithm converges to a local optimum. It is important to realize that the performance of k-means depends on the initialization of the cluster centers; a bad choice of initial seed, e.g. outliers or extremely close data points, can easily cause the algorithm to converge on less than globally optimal clusters. For this reason, it's usually a good idea to iterate k-means multiple times and choose the clustering that minimizes overall error.

2.5 Extraction of Frequent Phrases algorithm

To construct the document's bag of frequent words and phrases, we used a simple method based on Apriori algorithm.

Overview on Apriori

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori computes the frequent itemsets through several iterations known as a level-wise search. Each iteration has two steps: candidate generation and candidate counting and selection, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1.Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database [24].

Apriori algorithm [75] [76] to find frequent occurring phrases from a document collection or a transaction database. The Apriori algorithm consists of two steps: In the first step, it extracts frequent itemsets, or

phrases, from a set of transactions that satisfy a user-specified minimum support. In the second step, it generates rules from the discovered frequent itemsets. For this task, we only need the first step is shown in Figure2, i.e., finding frequent itemsets ({A} {B}{C} {E}{A C}{B C} {B E} {C E} {B C E}) that satisfy minimum support=2.

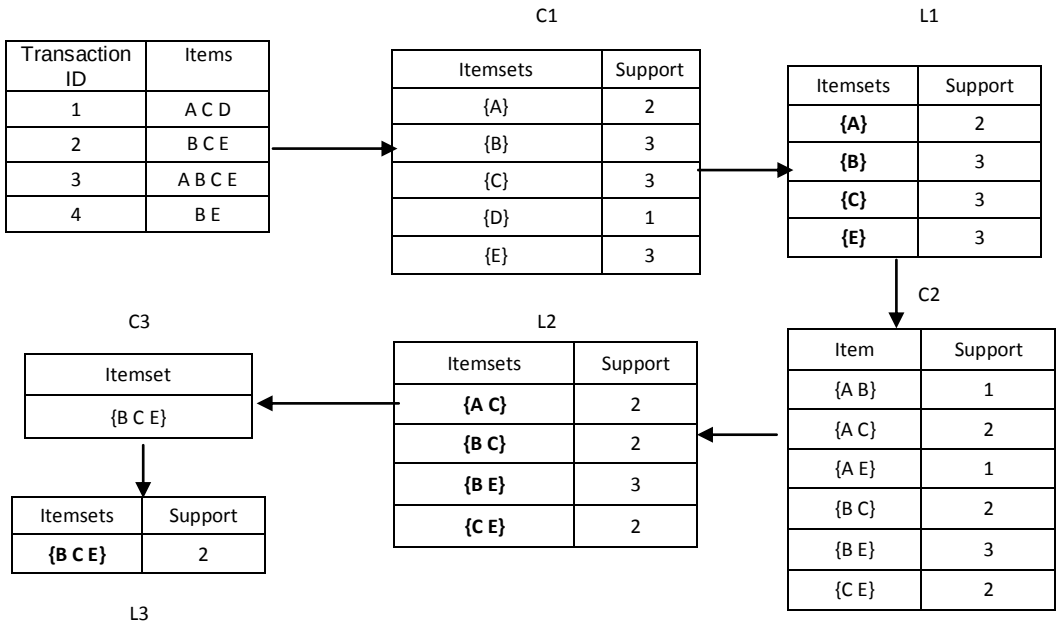

Figure 2.2 :Step 1 to find all frequent itemsets

2.6 Conclusion

This chapter introduced a definition of document clustering, different clustering algorithms, an overview of the most used techniques that deal with a document as a bag of words and techniques that used various semantic relations, ontology-based similarity measures, overview on similarity measures, some examples of Ontologies used to enhance document clustering. This forms an overall look at the process of document clustering and techniques to improve clustering results which is the top mission of this thesis.

Chapter 3

Related Work

This chapter reflects a number of researches that exploiting Background Knowledge in WordNet and Wikipedia to Enhance Document Clustering.

3.1 WordNet

Many studies have used WordNet as background knowledge to incorporate semantics into the bag of words to measure semantic similarity among words [40] [41] [35] [42] [20] [43] [44] [45].

A number of research efforts explored the use of WordNet as background knowledge to enhance document clustering by offering relations between vocabulary terms and the results have been different where some studies suggested that the use of an WordNet is helpful for clustering process, while others have reported that the WordNet is not helpful [46] [13] [47] [12] [16] [48] [14] [39].

The following researches used WordNet and they monitored the improvement in the results.

Hotho et al. [13] used WordNet synsets to augment document vector, showed that enhancing the bag of words with Wordnet synsets from the words in the text and their hypernyms (up to a certain distance) does make better clusters than a plain bag of words representation.

Recupero and Reforgiato [12], Wang and Hodges [14] used WordNet as background knowledge in document clustering with different datasets, The results have been showed that the use of an ontology is helpful for clustering

Other researches did not detect any improvement, a group of researchers concluded that WordNet does not benefit because its structure does not help in finding the similarity between the words.
Jing, L., et al. [15] used the same technique as Hotho et al. and enhances it by computing a word similarity measure based on what they call 'mutual information' over their clustering corpus. However, their technique didn't produce any considerable improvement over Hotho et al.'s baseline.

Passos and Wainer [49] showed that many similarity measures between words derived from Wordnet are worse than the baseline for the purposes of text clustering, Wordnet does not provide good word similarity data. Due to a variety of reasons the similarity between two words is not one of Wordnet's goals, and its structure does not fit well to the task, no measurements are directly based on Wordnet can relate a verb such as "to seat" to a noun such as chair.

Sedding and Kazakov [16] showed synonyms and hypernyms, disambiguated only by Part-of-Speech tags are not successful in improving clustering effectiveness. This could be attributed to the noise introduced by all incorrect senses that are retrieved from WordNet.

Fodeh et al. [46], Termier, A et al. [48] used WordNet with different datasets, The results have reported that the ontological concepts adds no value and sometimes impairs performance of document clustering.

Fodeh et al. [17] addressed the issue of the effect of incorporating the polysemous and synonymous into document clustering, that showed the polysemous and synonymous nouns play an important role in clustering,

even though their disambiguation does not necessarily lead to significant improvement in cluster purity.

Moravec et al. [47] showed different results when using two evaluation measures. Recall measure showed that, using wordnet improved clustering result. Whereas precision measure showed that, using wordnet did not improve clustering process.

3.2 Wikipedia

Techniques that employ Ontological features for clustering try to integrate the ontological background knowledge into the clustering algorithm. Ontology based similarity measures are often used in these techniques to calculate the semantic similarity between document terms. There is a plenty of Ontology-based similarity measures that exploit different ontological features, such as distance, information content and shared features, in order to quantify the mutual information between terms (reader is referred to [19] for a review and comparison of ontology based similarity measures). Distance between two document vectors is then computed based on the semantic similarity of their terms.

A number of research efforts explored the use of Wikipedia to enhance text mining tasks, including document clustering [64] [67] [68] , text classification [68]and information retrieval [69]. Few approaches have explored utilizing Wikipedia as a knowledge base for document clustering. [70] proposed and evaluated a method that is based on matching documents with the most relevant articles of Wikipeda, and then augmenting the document's BOW with the semantic features.

Spanakis et al. [71] proposed a method for conceptual hieratical clustering that exploits Wikipedia textual content and link structure to

create compact document representation. However, these efforts do not make use of the structural relations in Wikipeida. As a result, the semantic relatedness between words that are not synonyms is not considered when computing the similarity between documents.

Huang et al. [65] proposed an approach that maps terms within documents to Wikipedia's anchors vocabulary. Then they incorporated the semantic relatedness between concepts by using Milne and Witten measure [72] which takes into account all of the Wikipedi's hyperlinks. Our work is similar in that it also uses a similarity measure that is based on the Wikipedia's hyperlinks. However, their approach did not tackle the issue of frequent itemsets, and they instead used a less efficient approach by examining all possible n-grams. Another difference is the way the document similarity is measured: while they augmented the measure of document similarity, our approach augments the document's vector by reweighting the tf-idf score of each word according to its relatedness to other document's words. This makes our approach independent of, and can be used with, any measure of document similarity since the reweighting process is carried out before computing similarity between document pairs.

Another work that can be compared to ours is presented by Hu, X., et al [64]. They developed two approaches: exact match and relatedness-match, to map documents to Wikipedia concepts, and further to Wikipedia categories. Then documents are clustered based on a similarity metric which combines document content information, concept information as well as category information. However, their approach requires pre-processing of the whole Wikipedia's textual content, a thing that leads to substantial increase in both runtime and memory requirements. Instead, our approach does not require any access to the

Wikipedia's textual content, and relies only on the Wikipedia's link structure to compute similarity between terms.

Hu, J., et al. [73] proposed a method that mines synonym, hypernym and associative relations from Wikipedia, and append that to traditional text similarity measure to facilitate document clustering. However, their method was developed specifically for the task and has not been investigated independently. They also built their own method of measuring similarity through Wikipedia's category links and redirects. We instead used a similarity measure that is modeled after the Normalized Google Distance [66] which is a well-known and low-cost method of measuring similarity between terms based on the link structure of the Web.

Wikipedia has been employed in some efforts for short text classification. For example, Hu, X., et al. [67] proposed an approach that generates queries from short text and use them to retrieve accurate Wikipedia pages with the help of a search engine. Titles and links from the retrieved pages are then extracted to serve as additional features for clustering.

Phan, X et al. [74]presented a framework for building classifiers that deal with short text. They sought to expand the coverage of classifiers by topics coming for external knowledge base (e.g. Wikipedia) that do not exist in small training datasets. These approaches, however, use Wikipedia concepts without considering the hierarchical relationships and categories embedded in Wikipedia.

Chapter 4

Investigating The Influence Of WordNet on Document Clustering

4.1 Introduction

In most techniques of document clustering, documents are represented as bag of words, and then are assigned to clusters according to the similarity scores obtained from the cosine similarity measure. These techniques ignore the semantics between terms. As a result, a document that only contains the word “plane” and another that only contains the word “jet” are assigned to different clusters as the cosine similarity between them will be 0.

Existing research has tried to remove this limitation by proposing clustering techniques that are based on meanings similarities. The similarity between any two words can be measured either from an ontology or an electronic dictionary, or by inferring meaning from the context. Several efforts have investigated ways to integrate domain ontology as background knowledge in document clustering, and have shown that ontology semantics have the potential to improve the quality of the obtained clusters [15] [39] [13].

WordNet [35] is one of the most popularly used semantic networks for estimating semantic similarities. Wordnet has an ontology alike structure : words are represented as having several meanings (each such meaning forming a synset, which is the atomic structure of Wordnet), and relations between words (hyponymy, hyperonymy, antonymy, and other similar relations) are represented as links in a graph. Many similarity measures use the relations defined in WordNet to determine the semantic relatedness between words. Due to its wide coverage as compared to other restricted domain ontologies, many efforts used it as a background knowledge for document clustering. The similarity scores obtained from WordNet can be used to enhance the document’s representation by giving more weight to words that are semantically related. With the enhanced

document's representation, the clustering algorithm can better assign documents to clusters based on their semantic similarity to each other.

The purpose of this chapter is to explore the use of WordNet to enhance document clustering. We assume that the use of WordNet can help determine conceptual relationships between domain terms which do not match syntactically.

In our work, we will implement the experiment on a datasets which have been used in previous researchers as well as we will testing the experiment by using a new data set prepared from a group of IT experts.

Through our experiment we will evaluate the usage of semantic relations based on WordNet to enhance document clustering and compare our results with the results in previous researches.

4.2 Experimental Tools:

The following tools were used:

- Stanford Natural Language Processing toolkit for preprocessing steps.

The Natural Language Processing Group at Stanford University is a team of the faculty members, research scientists, postdocs, programmers and students who work together on algorithms that allow computers to process and understand human languages, Stanford CoreNLP contain a set of natural language analysis which can take raw English language text input and return the base forms of words including, tokenization, sentence splitting, the part-of-speech (POS) tagger, lemmatization.

- A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.
- A tokenization divides text into a sequence of tokens, which roughly correspond to "words".

- Lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form, for example am, are, is --> be ,car, cars, car's, cars' --> car.

The Stanford CoreNLP code is written in Java, for download it from http://nlp.stanford.edu/software/corenlp.shtml#Download

- WordNet is a lexical database for the English language that contain synonyms and records the various semantic relations between these synonyms.
- WS4J (WordNet Similarity for Java) provides a pure Java API for several semantic relatedness and similarity measures by download a jar file to use WS4J in java program, that used to measure similarity between terms based on WordNet ontology by multiple measures, download it from https://code.google.com/p/ws4j/

4.3 Experimental Design

We present the experimental setting that include Datasets, Conditions, Procedure, Evaluation Measures to evaluate the effectiveness of document clustering.

4.3.1 Experimental Settings:

In this section, we will describe the experimental environment of the experiments, and determine the experimental tools that are used in the experiments, final specify the setting of the experiments in the research.

4.3.1.1 Datasets

We conducted document clustering experiments in this chapter with three datasets: Reuters-21578, Journals and OHSUMED, the details of each dataset is given below.

Reuters-21578 [52] :

The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd**[52]**.

Reuters-21578 dataset has been widely used for evaluating document clustering algorithms, used in comparable studies before and freely available for download, but this dataset has several known limitations for example, some of documents are assigned to multiple classes or size of some categories is relatively large while others have few documents [17], Its domain is not specific, therefore it can be understood by a non-expert[16]. Therefore, we sampled a dataset that contains the 100 documents from 5 labeled classes.

Journals :

We have collected 100 abstracts from international journals such that these journals have classified to five different topics, data mining, Software Engineering, Human Computer Interaction, Software Quality Assurance, Semantic Web, through two experts in the Information Technology

We started by creating a journal dataset under all conditions which implemented on Reuters dataset in order to clarify the results. Through different results among previous studies on Reuters dataset, we conclude

that data set could be the reason of the difference in the results in the experiment on Reuters dataset.

The motivation to create this data set is to have files that have strong semantic relations in between. The journal files are all related to computer science topics and often use similar or related vocabulary.

Whereas the files in the Reuters dataset, which contain news of different categories, often use uncorrelated words and numbers. We aim to explore if the type of the data set could affect of the value of the background knowledge. Our hypothesis is that applying WordNet-based similarity measures on Reuters dataset may not improve the clustering results due to the diversity of information content.

To sum up, we have chosen two datasets, the first dataset is widespread and found in most previous studies, and the second dataset is new dataset to conduct the experiment not on traditional datasets, the dataset of abstract can be download it from the URL: gate.alazhar.edu.ps/ datajournals.rar

OHSUMED[53]:

The OHSUMED test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The National Library of Medicine has agreed to make the MEDLINE references in the test database available for experimentation[54]. We sampled a dataset that contains the 100 documents from 5 labeled classes.

4.3.1.2 Conditions

The experiment contains several tests under specific conditions,

1- Traditional clustering without background knowledge: This is the baseline case.

2- Enhance the document representation by identifying and replacing synonyms: Documents may use different synonyms of the same term. Without identifying synonyms, the classifier will treat synonyms as they are different words and may map them to different clusters. To resolve this issue, WordNet is used to identify synonyms and then replace them with a single term, e.g. a single term to represent all synonyms. Thus, each term will have a unique representation across different documents regardless of the different synonym words. By using a single word instead of different synonym word of the same term, that word will gain more weight in the document's vector representation.

3- Enhance the document representation by integrating semantic relatedness between terms: The previous setting aims to enhance the document representation by replacing only the synonyms, but it does not consider the semantic relatedness between other terms. Terms that are not synonyms can be semantically related, (e.g. desert, camel). In this test setting, the WordNet is used not only to identify synonyms, but also to measure the degree of similarity between terms, and then integrate the similarity scores into the document representation. The idea is that terms should gain more weights according to its semantic relatedness to other terms in the document. Different similarity measures are used and assessed in this test.

4.3.1.3 Procedure

- Document Preprocessing: This step is used in all the tests and with all datasets to transform documents that contain strings of characters into a suitable representation for the clustering task, that include several preprocessing steps:

a- **Tokenization** is the process of splitting sentences into individual tokens, which roughly correspond to "words". These tokens becomes input for further processing.

b- **Stop-words Removal**: The stop-words are high frequent words that carry no information, stop-words are filtered out based on group of words which can be chosen as the stop words, e.g. pronouns, prepositions, conjunction, numbers.

c- **Stemming**: By word stemming through group of words that carry the same conceptual meaning, such as connected, connect, connection, we used the light stemming instead of the ordinary stemming. Light stemming (or lemmatization) preserves the root of the word as found in the dictionary.

- Implementation of K-means algorithm

In our experiment, we implemented K-means algorithm in Java and used it for our experiment. Although some platforms and tools such as RapidMiner or Weka offer ready-made solutions for document clustering without having to implement the clustering method itself, these solutions use common settings and they often follow predefined steps which cannot be easily altered to cope with our experimental needs. Building our own implementation of K-means allows us to easily interfere in the steps of data processing and clustering to apply our testing condition by, for example, incorporating the semantic scores obtained from WordNet.

It should be noticed that the approach of enhancing the document representation by exploiting WordNet semantics is performed independently of the clustering algorithm. Only the document representation is augmented with similarity measures. The procedure of clustering algorithm is not modified but it is expected to produce better results. Therefore, any traditional clustering algorithm can be used to assess the value of enhancing the document representation on the clustering results. Different Testing conditions will result in different document representation but the clustering algorithm remains intact to avoid any biased results. In our experiment, K-means was used across the testing conditions because it easy to implement and is widely used by similar studies from the state of the art. This allows making our results comparable with other approaches from the state of the art which also used K-means for clustering.

In our experience we set K = "5" since each dataset consists of five labelled classes. Since the clustering results of K-means is influenced by the initial selection of cluster centroids, for each evaluation based on K-means, we run ten times with ten random initialization and take the average as the final clustering result. For the comparative experiment, we used the same initialization of result in other tests.

In the following subsections, each test is explained in detail:

Test 1: Traditional Document Clustering

Traditionally, document representation is based on the use of the bag of words.

Most of the document clustering methods are based on the Vector Space Model which is widely used as data model for classification and clustering[28]. Documents are represented using the vector space model (VSM). This model is known as term frequency-inverse document

frequency model (tf-idf)[55], which ranks the importance of a term in its contextual document corpus. The steps in test 1 (as in Figure1).

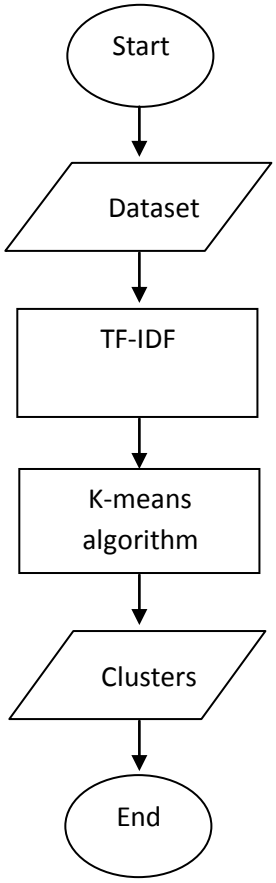

Figure 4.1: Flow Chart-Traditional Document Clustering

Given a document set D, $T = \{t_1, t_2, t_3, .., t_n\}$ is the set of terms in D. Then, a document $d_i \in$ D can be represented as a term dimension vector

$$d_i = tfidf(d_i, t_1), tfidf(d_i, t_2), .. \quad (1)$$

Where $tfidf(d_i, t_1)$ is a weighting of term t_n in document d_i. The $tfidf$ weighting can be defined as follows.

$$tfidf(di, tn) = tf(di, tn) * \log\left(\frac{|D|}{df(tn)}\right) \quad (2)$$

Where $tf(di, tn)$ is the frequency of term tn in document d_i; $df(tn)$ is the document frequency that indicates the number of documents

containing tem t_n. After representing each document in tf-idf model, traditional K-means algorithm is applied.

Note that this approach ignores the conceptual relations between terms, and weighs terms only according to their frequency of co-occurance in the document collection.

Test 2: Enhancing the Document Representation by Replacing Synonyms

One limitations of using vector space model is that different vector positions may be allocated to the synonyms of the same term. For example, the terms {smart, brilliant, bright} are weighted separately although they are all synonyms. This leads to information loss because of the importance of a determinate concept is distributed among different vector components. Previous studies (e.g.[56]) approached this issue by referring to lexical databases like WordNet to identify synonyms , or synsets, and reweigh them accordingly. Similarly, our approach will refer to WordNet in order identify synonyms of a particular concept, assuming that the ontology is properly populated with all synonyms of domain concepts.

After identifying all synonyms of a single term in the document collection, the document's bad of words will be modified by replacing all synonyms with a single descriptor, i.e. representing term. For example, the terms {smart, brilliant, bright} will be replaced by a single term {intelligent}. Afterwards, the document is represented by using the tf-idf model. Therefore, the representing term will have a cumulative weight that is equal to the sum of tf-idf weights of replaced synonyms. Finally, K-means clustering algorithm is applied.

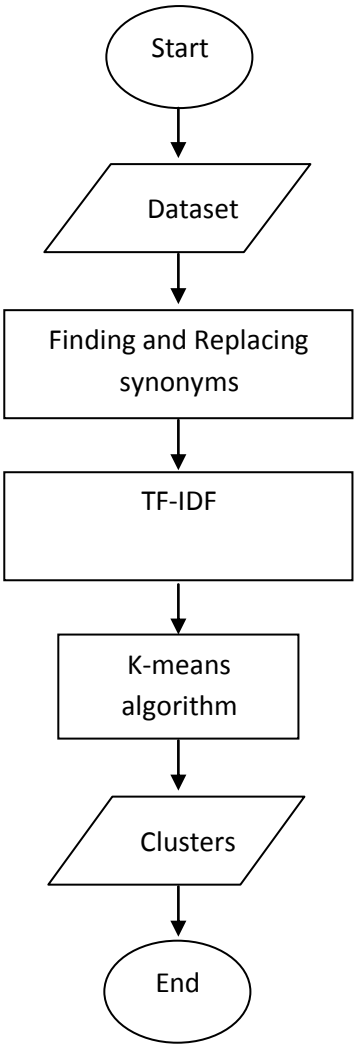

Figure 4.2: Flow Chart- Enhancing the Document Representation by Replacing Synonyms

The steps of implementation of test2 (as in Figure2)

Before finding synonyms in documents, our test follows the following preprocessing steps,

First, all documents are broken down into sentences., these sSentences are then undergone part of speech tagging (Standford POS tagger was used). Part of speech tags are required by WordNet to identify synonyms words as it assumes that synonyms should have the same POS tag.

After tagging the content of documents, other preprocessing steps including tokenization, stopword removal and light stemming are

applied. Note that these preprocessing steps cannot be applied prior to POS tagging which requires original text.

The next step is to search the documents for terms that are synonyms with the help of WordNet. Synonyms of a particular concept are all replaced with a representing term in the documents' bag of word.

After replacing synonyms, documents are represented using the term vector model with tf-idf weighting. Finally, the K-means algorithm is applied.

Test 3 : Enhancing the document representation with semantic similarity scores obtained from WordNet.

Having documents with different terms sets does not necessarily mean that documents are unrelated. Document terms can be semantically related even though they are syntactically different. For example the terms {Gaza strip, Palestine, Jerusalem, Mediterranean sea} are all related with some relationships which cannot be captured without using a background knowledge.

In the previous test (Test 2), we sought to enhance the document's representation by identifying and replacing only synonyms of the same term. Terms that are semantically related and that are not synonyms are still not considered. For example, the similarity between the two words: <camel> and <desert> is not ignored by measuring similarity between documents.

Test 3 aims to overcome this limitation by representing the document in a way that reflects the similarity in meanings of the document's terms. Common similarity measures (refer to section 2.3.1) are used to measure

the similarity between each pair of the document terms, and then similarity scores are incorporated into the docment's representation. Similarity measures exploit knowledge retrieved from a semantic network (i.e., WordNet) to measure the degree of similarity between terms.
Similarity measures use different algorithms to define the topological similarity, by using the WordNet ontological structure, to define the distance between terms. For example, some measures (e.g. Leacock and Chodorow, 1998) relies on the shortest ontological path between terms for their measure of similarity.

In our experiment, ontology-based similarity measures are used to estimate the similarity scores between term pairs according to the topology structure of WordNet. These similarity scores are then incorporated into the document's vector representation so that terms are semantically related will gain more weight. Reweighting terms according to their semantic elatedness may help discount the effects of class-independent general terms and aggravate the effects of class-specific "core" terms[19]. This can eventually help to cluster documents based on their meaning. In addition, we examined the use of different similarity measures in order to explore best similarity measures to use with WordNet.

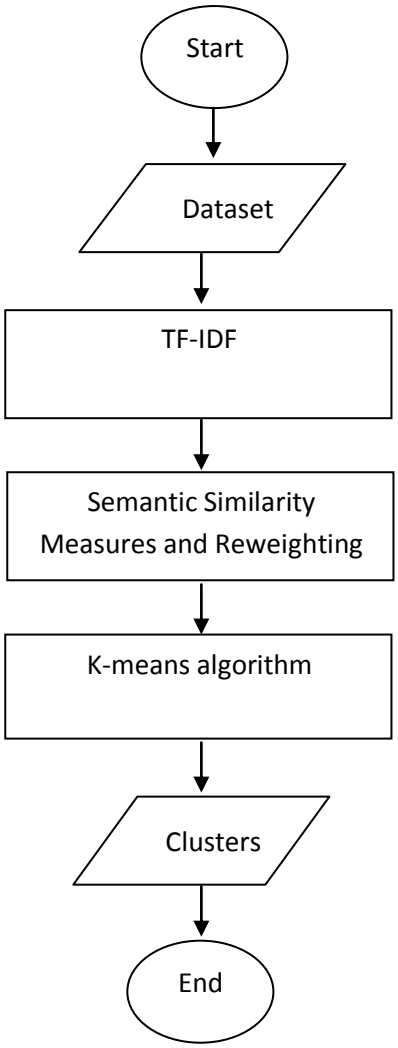

Figure 4.3: Flow Chart- Enhancing with semantic similarity scores obtained from WordNet.

This test was done through the following steps (as in Figure3):

1. Preprocessing step which consists of tokenization, stop-word removal and stemming.
2. $d = \{w_1, w_2, w_3, .., w_n\}$ be the document's vector representation.
   - where w_i is the weight of term t_i in document d, and is computed using the term frequency - inverse document frequency ($tf.idf$) model.
3. The semantic similarity between each pair of terms in the document's bag of words is calculated by using each similarity measure shown in Table 2.1.
4. The weights of terms will be adjusted using the following equation[58]:

$$w'_i = w_i + \sum_{j=0,j\#i}^{m} w_j * \text{sim(i, j)} \qquad (3)$$

- Where: w'_i stands for the augmented term weight of term i, w_i is the weight of term i computed with the $tfidf$ model, $w_j$is the weight of term j of the same document, and $sim(i,j)$ is the semantic similarity between terms i, j which rates between 0 and 1.
- This equation will result assigning higher weights to semantically related terms within the set of document terms.
- The weights of terms that are not semantically related to any other terms or that are not mapped to any ontology concepts will remain unchanged.

5. Then, K-means is applied on the augmented VSMs same as in Test1.

The steps from 3 to 5 are repeated from every similarity measure in Table.

4.3.1.4 Evaluation Measures

We evaluate the effectiveness of the document clustering by two quality measures F-measure[50], purity[51].

F-Measure

The F-measure uses a combination of precision and recall values of clusters. We let n_i designate the number of documents in class i, and c_j designate the number of documents in cluster j. Moreover, we let c_{ij} designate the number of items of class i present in cluster j. Then we can

define *prec(i, j)*, the precision of cluster j with respect to class i and *rec(i,j)*, the recall of a cluster j with respect to class i

as $prec(i,j) = \frac{C_{ij}}{C_j}$ and $rec(i,j) = \frac{C_{ij}}{n_i}$. The f-measure, *F(i,j)*, of a class i with respect to cluster j is then defined as

$$F(i,j) = \frac{2 * prec(i,j) * rec(i,j)}{prec(i,j) + rec(i,j)}$$

The f-measure for the entire clustering result is defined as

$$F = \sum \frac{n_i}{n} \max\left(F(i,j)\right)$$

Purity

Purity measures the dominance of the largest class per cluster, it can be defined as the maximal precision value for each class j, We compute the purity for a cluster j as $purity(j) = \frac{1}{C_j} \max\left(C_{ij}\right)$. We then define the purity of the entire clustering result as:

$$purity = \sum \frac{C_j}{N} purity(j)$$

Where $N = \sum_j C_j$, i.e. the sum of the cardinalities of each cluster, Note that we use this quantity rather than the size of the document collection for computing the purity.

For Purity and F-measure ranging from 0 to 1, the bigger the value is the higher quality the clustering has[19].

4.4 Results and Discussion

Table 4.1 shows the clustering results in terms of Purity and F-measure. The rows of the table depict the three experimental tests we conducted and which include:

Test1 (Without Semantics): this is the baseline case which involves applying the clustering method. i.e. K-means, without exploiting background knowledge.

Test2 (With Synonyms): This test uses the WordNet to identify and replace the synsets, or synonyms, of each term with a unique descriptive term. Grouping all synonyms as a one concept in WordNet has an effect on increasing or decreasing the semantic similarity between documents, which in turn affects document clustering.

Test3 (With Similarity scores): This test uses the similarity scores obtained from a variety of similarity measures to incorporate meanings in the document's representation. Terms are reweighted to have more or less weights according to their similarity to other terms in the document.

Note that for this test, different similarity results were used, but only the best similarity scores are shown in this table.

The columns of Table 4.1 depicts the three datasets which we run our experiment over.

	Reuters		Journals		OHSUMED	
	Purity	F-measure	Purity	F-measure	Purity	F-measure
Test1: Without Semantics	0.57	0.64	0.60	0.79	0.36	0.47
Test2: With Synonyms	0.64	0.77	0.80	0.95	0.49	0.6
Test3: Similarity scores	0.59 (LCH)	0.70 (LCH)	0.68 (WUP)	0.86 (WUP)	0.43 (RES)	0.65 (RES)

Table 4.1:The results for different experimental tests on datasets.

Comparing the clustering results from the different datasets, we noticed that:

Using the Reuters dataset, the clustering with synonyms (Test 2) gave the best results (F-score =0.77 and purity =0.64), followed by the

clustering with similarity scores (Test 3) (F-score=0.70, Purity=0.59), followed by the clustering without any semantics (F-score=0.64, Purity=0.57).

This result indicates that incorporating semantics by replacing synonyms with WordNet concepts has the best impact on the clustering results. On the other hand, the use of similarity measures [test 3] has unexpectedly produced results that are slightly better than the baseline case (clustering without semantics) but worse than results obtained from the clustering with synonyms. The lower performance of the incorporated similarity measures can be explained by the noise they caused to the document's representation vector that ended up producing close to the baseline case.

Using Journal and OHSUMED datasets, it was obvious that the clustering with synonyms has also produced better clustering results followed by the clustering with similarity measures. Again, this proves that the use of WordNet has improved the clustering results as compared to clustering without semantics.

Comparing the results obtained from the three datasets, we can see that the improvement resulted from semantic-based approaches (synonyms and similarity measures) was obvious in the case of Journals and OHSUMED datasets than in the case of Reuters datasets. This difference can be explained by the nature of the dataset which can sometimes hinder the ability to measure similarity between terms. For example, the Reuters dataset is heterogeneous in nature and includes content related to different domains and news. It is often difficult to identify semantic relations between terms related to different domains. Therefore, WordNet had little impact on the obtained clusters in case of the Reuters dataset.

However, The Journals and OHSUMED datasets are domain-specific, a thing that makes it easy to identify terms that belong to a specific domain and measure similarities between them. This explains the better results

obtained in these cases as compared to the results obtained from Reuters dataset.

The above discussion reveals that the use of different datasets can result in different clustering results: The more homogeneous and domain-specific the dataset is, the easier it becomes to capture similarities between terms included in the dataset, and hence the more influence the WordNet has on the clustering results.

We should also bear in mind that the WordNet is a general-purpose lexical database of English terms but it does not provide a thorough coverage of specific domains of knowledge. Although its use has improved the clustering performance in our experiment, WordNet is not meant to be used with domain specific applications. Therefore, using WordNet for clustering domain-specific datasets is unlikely to produce significant semantic enhancements in all cases. It is always recommend to use domain-specific ontologies to cover domain-specific datasets

Results also indicated the use of similarity measures for clustering has not produced the best results as expected, and the improvement resulted from using them was always less that the improvement resulted by replacing synonyms. This result conforms to some previous efforts which indicated that the similarity measures have little impact on the clustering results and may even produce worse results. e.g. Jing, L., et al. [31], Passos and Wainer [47]. In particular, using similarity measures with WordNet may produce noise that can hinder the document representation and in turn disrupt the clustering results.

This result also shows that using similarity measures with WordNet does not seem to improve the clustering results. This can be attributed to the structure of WordNet which is mainly designed to represent specific relations (e.g. hyponymy, hyperonymy) but is not designed to capture

similarity between words. For example, when measuring the similarity between the words: “camel” and “desert”, or between the verb “sit” and the noun “chair”, the similarity scores were close to 0.

Table 4.2 list the different similarity measures we used for test 3 (With similarity scores) and the clustering performance per each measure.

Similarity Measures	Rueters		Journals		OHSUMED	
	Purity	**F-measure**	**Purity**	**F-measure**	**Purity**	**F-measure**
PATH	0.57	**0.68**	**0.64**	**0.83**	**0.38**	**0.5**
LCH	**0.59**	**0.70**	0.60	**0.7**	**0.39**	**0.49**
WUP	0.56	0.64	**0.68**	**0.86**	**0.41**	**0.55**
JCN	0.40	0.48	0.43	0.5	0.30	0.39
LIN	0.48	0.55	**0.65**	0.79	**0.41**	**0.55**
RES	0.48	0.61	**0.63**	0.77	**0.43**	**0.65**
LESK	0.54	**0.67**	0.58	0.69	**0.42**	**0.57**
HSO	0.46	0.58	0.56	**0.84**	**0.42**	**0.62**

Table 4.2:The results for different similarity measures on datasets.

Comparing the use of different similarity measures, result also vary: in the case of Reuters datasets, the LCH measure gave the best results followed by the PATH and WUP measures. When using the Journals dataset, the WUP measure was the best one, followed by the PATH and LIN measures. In the case of OHSUMED, the RES measure gave the highest results, followed by the HSO and LESK. However, the improvement on the results was not significant [t-test, $p>0.05$].

These results indicate that there was no certain measure to ensure the best clustering results.

4.5 Conclusion

In this chapter, we evaluate WordNet as a background knowledge in improving document clustering process by using synonyms and semantic similarity measures between terms through different experimental conditions. These conditions involve the use of different datasets,

different similarity measures and different preprocessing steps to resolve a conflict regarding the value of WordNet as background knowledge for document clustering showed in previous studies and answer what factors could make WordNet useful in particular situations and while not in others situations and do the different datasets have impact on the clustering results.

This chapter contains several tests under specific conditions, the first test traditional clustering without background knowledge, the second test by identifying and replacing synonyms and third test by integrating semantic relatedness between terms using different measures and we implemented K-means algorithm in Java and used it for our tests.

The result indicates that incorporating semantics by replacing synonyms with WordNet concepts has the best impact on the clustering results.

Whereas the use of similarity measures has unexpectedly produced results that are slightly better than the traditional clustering and worse than results obtained from the clustering with synonyms.

This result supports many previous efforts which indicated that the similarity measures have little impact on the clustering results and WordNet does not provide good word similarity data.

Chapter 5

An Efficient Approach for Semantically-Enhanced Document Clustering by Using Wikipedia Link Structure

5.1. Introduction

Some approaches have used Wikipedia concepts and category information to enrich document representation and handle the semantic relationships between document terms [64, 65]. Wikipedia is much more comprehensive than other ontologies since it capture a wide range of domains, is frequently updated and well structured. Wikipedia can be seen as an ontology where each article represents a single ontology concept, and all concepts are linked together by hyperlinks. In addition, Wikipedia has a hierarchal categorization that resembles the structure of an ontology whereas each article belongs to one or more information categories.

In this chapter, we propose an approach to improve document clustering by explicitly incorporating the semantic similarity between Wikipedia concepts into the document's vector space model. Our approach is distinguished over similar approaches in terms of the way we used to efficiently map the document content to Wikipedia concepts and the low-cost measure we adapted to determine semantic similarity between terms. In the following section, we discuss similar efforts that also exploited knowledge from Wikipedia to enhance document clustering, and compare their approaches with ours.

5.2. An Approach For Wikipedia-Based Document Clustering

The pseudo code of our approach for Wikipedia-based document clustering is shown in Figure 4.1, and consists of three phases: The first phase includes of a set of text processing steps for the purpose of determining terms that best represent the document content. In the second phase, each document is represented by using the tf-idf weighted vector. Document terms are then mapped to Wikipedia concepts. In the third phase, the similarity between each pair of Wikipedia concepts is

measured by using the Wikipedia link structure. The tf-idf weights of original terms are then reweighted to incorporate the similarity scores obtained from Wikipedia. By the end of the algorithm, the tf-idf representation of each document is enriched so that terms that are semantically related gain more weight. Documents can then be clustered using any traditional clustering method such as k-means. These phases are explained in detail in the subsequent sections.

Prior to applying our approach, Wikipedia's vocabulary of anchor text is retrieved from the Wikipedia dump, which is a copy of all Wikipedia content, and stemmed in order to be comparable with the stemmed

Input: A set of documents $D = \{d_1, .., d_n\}$
Begin
{**Phase 1**: Pre-processing and extraction of frequent phrases}
for each document *d* ∈ *D* do
  Apply Tokenization, stemming and stopword removal
end for
Concatenate all documents
Apply Apriori algorithm to extract frequent itemsets

{**Phase 2**: Construct tf-idf weighted vectors and map terms to Wikipedia concepts}
for each document *d* ∈ *D* do
  Discard tokens that overlap with frequent phrases
  Discard rare terms
 Build the BOW of *d* where BOW = Retained tokens ∪ frequent phrases
  for each term *t* ∈ *BOW* do
    Calculate *tf-idf* for *t*
    if *t* matches Wikipedia concept(s) then
      Replace *t* with matching Wikipedia concept(s)
    end if
  end for
end for

{Phase 3: {Reweighting *tf-idf* weights}
for each document *d* ∈ *D* do
  for each term t_i ∈ BOW of *d* do
    for each term t_j ∈ BOW of *d* AND $t_i \neq t_j$ do
      Compute similarity between $t_i and\ t_j$ using equation 1
      if $t_i or\ t_j$ are ambiguous then
        Perform word-sense disambiguation (see section 6)
      end if
    end for
    Reweight $tdidf(d, t_i)$ using equation 2
  end for
end for

{Document clustering}
Apply any conventional clustering algorithm (e.g. k-means)
End

Figure 5.1:Pseudo code of our algorithm of document clustering

document content. For measuring similarity between Wikipedia concepts,

all outgoing hyperlinks, incoming hyperlinks and categories of articles are also retrieved. Note that this task incurs a one-time cost, thus allowing the clustering algorithm to be invoked multiple times without the additional overhead of reprocessing the Wikipedia content.

5.3. Construction of Document's Vector Space Model

The first step of our clustering approach is to represent each document as a bag of words (BOW). Note that traditional clustering algorithms treat a document as a set of single words, thus losing valuable information about the meaning of terms. When incorporating semantics in document clustering, it is necessary to preserve phrases, the consecutive words that stand together as a conceptual unit. Without preserving phrases, actual meanings of document terms may be lost, making it difficult to measure semantic similarity in between. For example, the phrase "big bang theory" refers to a concept that is entirely different from what its individual tokens refer to. Thus, we aim to create a document's BOW representation whose attributes include not only single words but also phrases that have standalone meanings. This phase starts with some standard text processing operations including stopword removal and word stemming. Stopwords are words that occur frequently in documents and have little informational meanings. Stemming finds the root form of a word by removing its suffix. In the context of mapping with Wikipedia concepts, stemming allows to recognize and deal with variations of the same word as if they were the same, hence detecting mappings between words with the same stem.

We used a simple method based on Apriori algorithm to find frequent occurring phrases from documents. A phrase is defined as frequent if it

appears in n number of documents (For our task we set n = 3). In addition, the algorithm was restricted to find itemsets with four words or fewer as we believe that most Wikipedia concepts contain no more than four words (this restriction can be easily relaxed).

After extracting frequent itemsets, we perform word tokenization to break the document text into single words. Many of the resulting tokens can be already part of the extracted itemsets. Therefore, we remove tokens that overlap with any of the retrieved frequent itemsets. Stemmed tokens as well as frequent itemsets that occur in the document will be combined together to form the BOW representing the document. Rare terms that infrequently occur in the document collection can introduce noise and degrade performance. Thus, terms that occur in the document collection less than or equal to a predefined threshold are discarded from the document's BOW.

It is worth noting here that similar works that exploited Wikipedia for document clustering often did not consider mining frequent itemsets occurring in the document [64] [65]. Instead, they extract all possible n-grams from the document by using a sliding widow approach and match them with the Wikipedia content. In contrast, our approach of extracting frequent itemset prior to the concept-mapping process is more time-efficient as it avoids the bottleneck of matching all possible n-grams to Wikipedia concepts.

After identifying the document's BOW, the next step is to map terms within the BOW to Wikipedia concepts: Each term is compared with Wikipedia anchors, and matching terms are replaced by the corresponding Wikipedia concepts. Terms that do not match any

Wikipedia concept are not discarded from the BOW in order to avoid any noise or information loss.

Formally, let $D = \{d_1, \dots, d_n\}$ be a set of documents and $T = \{t_1, \dots, t_m\}$ be the set of different terms occurring in a document . Note that T includes both: 1) Wikipedia concepts which replace original terms after the mapping process. 2) Terms that do not match any Wikipedia concepts. The weight of each document term is then calculated using tf-idf (term frequency-inverted document frequency). Tf-idf weighs the frequency of a term in a document with a factor that discounts its importance when it appears in almost all documents. The tf-idf of term t in document d is calculated using the following equation:

$$\text{tfidf}(d, t) = \log(\text{tf}(d, t) + 1) * \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

The document's vector representation $\overrightarrow{t_d}$ is then constructed from the tf-idf weights of its terms:

$$\overrightarrow{t_d} = \left(\text{tfidf}(d, t_1), \dots \text{tfidf}(d, t_m)\right)$$

5.4. Measuring Semantic Similarity Between Wikipedia Terms

After representing the document as a vector of term tf-idf weights, the next step is to augment these weights so that terms gain more importance according to their semantic similarity to the other document terms.

To measure similarity between document terms, we used a measure that is based on the Normalized Google Distance Measure (NGD)[66]. This measure is a relative semantic distance relies on the World Wide Web and a search engine such as Google or any other large electronic database, for instance Wikipedia that returns aggregate page counts to find out the similarity metric between terms. The NGD measure first uses

the Google search engine to obtain all Web pages mentioning these terms. Pages that mention both terms indicate relatedness, while pages that mention only one term indicate the opposite. The NGD denotes the distance or dissimilarity between two terms: the smaller the value of NGD, the more related the terms are semantically.

By using the Normalized Google Distance(NGD), is defined below, one can find the similarity between terms (0 for identical and 1 for unrelated).

$$NGD\ (x,y) = \frac{\max\{\log\ f(x),\log\ f(y)\} - \log f\ (x,y)}{\log N - \min\{\log\ f(x),\log\ f(y)\}}$$

Where x is the term1, y is the term2, $f(x)$ denotes the search results count of x, $f(y)$ denotes the search results count of y, $f(x,y)$ denotes the search results count of (x,y), and N is Total no. of pages searched by the search engine

For this work, the measure is adapted to exploit Wikipedia articles instead of the Google's search results. Formally, the Wikipedia-based similarity measure is:

$$sim\ (s,t) = 1 - \frac{\max\{\log(S),\log(T)\} - \log\ (S \cap T)}{\log(R) - \min\{\log(S),\log(T)\}} \quad (1)$$

where s and t are a pair of Wikipedia concepts. S and T are the sets of all Wikipedia articles that link to s and t respectively, and R is set of all Wikipedia concepts. The output of this measure ranges between 0 and 1, where values close to 1 denote related terms while values close to 0 denote the opposite. Note that the advantage of this measure is its low computational cost since it only considers the links between Wikipedia articles to define similarity.

After computing the similarity between each pair of terms, the tf-idf weight of each term is adjusted to consider its relatedness to other terms

within the document's vector representation. The adjusted weight w of a term t is calculated using the following equation:

$$\mathrm{w}(d,t_i) = tdidf\,(d,t_i) + \sum^{N}_{\substack{j=0, j\neq i \\ sim(t_i,t_j)\geq threshold}} tdidf\left(\mathrm{d},\mathrm{t}_j\right) * sim(t_i,t_j) \qquad (2)$$

where $sim(t_i, t_j)$is the semantic similarity between the terms t_i and t_j, and is calculated using Equation 1. N is the number of co-occurred terms in document d. The threshold denotes the minimum similarity score between two terms. Since we are interested in emphasizing more weight on terms that are more semantically related, it is necessary to set up a threshold value.

Note that this measure assigns an additional weight to the original term's weight based on its similarity to other terms in the document. The term weight remains unchanged if it is not related to any other term in the document or if it is not mapped to any Wikipedia concept. The final document's vector $\overrightarrow{t_d}$ is:

$$\overrightarrow{t_d} = \left(w(d,t_1), \ldots w(d,t_m)\right)$$

After constructing the semantically-augmented vectors for all documents, any conventional measure of document similarity, such as the cosine measure, can be used to measure similarity between document pairs. Note that in our approach we incorporate the similarity scores in the document representation before applying the document similarity measure. Thus, our approach is independent of, and hence can be used with, any similarity measure or clustering algorithm.

5.5. Word Sense Disambiguation

Concepts mentioned in Wikipedia are explicitly linked to their corresponding articles through anchors. These anchors can be considered

as sense annotations for Wikipedia concepts. Ambiguous words such as “eclipse” are linked to different Wikipedia articles based on their meanings in the context where they occur (e.g. eclipse "astronomical event", eclipse "software suite", eclipse "foundation"). When mapping document terms to Wikipedia concepts, it is necessary to perform word sense disambiguation to identify the correct word sense. Failing to do so may result in false results when measuring similarity between terms.

One way to disambiguate words is to simply use the most common sense. The commonness of a sense is identified by the number of anchors that link to it in Wikipedia. For example, over 95% of anchors labelled as “Paris” link to the capital of France while the rest link to other places, people or even music. However, choosing the most common sense is not enough and it is not always the best decision. Instead, we used the same approach used in [66]which uses the two terms involved in the similarity measure to disambiguate each other. This is done by selecting the two candidate senses that most closely related to each other. We start by choosing the top common senses for each term (For simplicity, we ignore senses that contribute with less than 1% of the anchor’s links). We then measure the similarity between every pair of senses, and the two senses with the highest similarity score are considered.

5.6. Evaluation

Wikipedia releases its database dumps periodically, which can be downloaded from http://download.wikipedia.org. The Wikepedia dump used in this evaluation was released on the 13th August 2014, and contains 12100939 articles. The data was presented in XML format. We used the WikipediaMiner [77] toolkit to process the data and extract the categories and outlinks out of Wikipedia dump.

5.6.1. Methodology

Our objective was to compare our approach with other approaches from the state of the art. Therefore, we used the same evaluation settings used by [73] in order to make our results comparable with theirs. The following two test sets were created:

- Reuters-21578 contains short news articles. The subset created consists of categories in the original Reuters dataset that have at least 20 and at most 200 documents. This results in 1658 documents and 30 categories in total.
- OHSUMed contains 23 categories and 18302 documents. Each document is the concatenation of title and abstract of a medical science paper.

Besides our method, we implemented and tested three different text representation methods, as defined below:

- Bag of Words: The traditional BOW method with no semantics. This is the baseline case.
- Hotho et al.'s method: this is a reimplementation of Hotho et al.'s WordNet-based algorithm [13]. The intention of considering this method is to compare how the use of Wikipedia as background knowledge influences the clustering results as compared to WordNet.

To focus our investigation on the representation rather than the clustering method, the standard k-means clustering algorithm was used. We used two evaluation metrics: Purity and F-score. Purity assumes that all samples of a cluster are predicted to be members of the actual dominant class for that cluster. F-score combines the information of precision and recall which is extensively applied in information retrieval.

5.6.2. Results

Table 5.1 shows how the different methods perform in clustering on the two datasets. In general, the performance of BOW on both datasets is improved by incorporating background knowledge either from WordNet (Hotho et al.'s method) or Wikipedia (our method). For instance, according to the F-score, for the Reuters dataset, our method and Hotho et al's method achieve 31% and 9% respectively.

On comparing the use of Wikipedia to WordNet, our approach outperformed the Hotho et al.'s approach for both datasets. Our approach achieves the best F-score and purity on both datasets. We applied t-test to compare between the performance of our approach and the others. Results show that our approach significantly outperformed other methods on the Reuters dataset with the p-value < 0.05.This demonstrates the potential of integrating Wikipedia as a knowledge source as compared to the WordNet based method.

	Reuters 21578		OHSUMed	
	Purity (Impr.)	F-score (Impr.)	Purity (Impr.)	F-score (Impr.)
Bag of Words	**0.57**	**0.64**	**0.36**	**0.47**
Hotho et al.	**0.59 (4%)**	**0.70 (9%)**	**0.39 (8%)**	**0.49 (4%)**
Our Approach	**0.73 (28%)**	**0.84 (31%)**	**0.52 (44%)**	**0.60 (27%)**

Table 5.1. Comparison with related work in terms of purity and F-score

5.7. Conclusion

Traditional techniques of document clustering do not consider the semantic relationships between words when assigning documents to clusters. For instance, if two documents talking about the same topic do that using different words (which may be synonyms or semantically associated), these techniques may assign documents to different clusters.

Previous research has approached this problem by enriching the document representation with the background knowledge in an ontology. This chapter presents a new approach to enhance document clustering by exploiting the semantic knowledge contained in Wikipedia. We first map terms within documents to their corresponding Wikipedia concepts. Then, similarity between each pair of terms is calculated by using the Wikipedia's link structure. The document's vector representation is then adjusted so that terms that are semantically related gain more weight. Our approach differs from related efforts in two aspects: first, unlink others who built their own methods of measuring similarity through the Wikipedia categories; our approach uses a similarity measure that is modelled after the Normalized Google Distance which is a well-known and low-cost method of measuring term similarity. Second, it is more time efficient as it applies an algorithm for phrase extraction from documents prior to matching terms with Wikipedia. Our approach was evaluated by being compared with different methods from the state of the art on two different datasets. Empirical results showed that our approach improved the clustering results as compared to other approaches.

Chapter 6

Conclusions and Future Work

In this research, we investigated approaches of document clustering enhancement by exploiting background knowledge such as WordNet and Wikipedia.

In the first part, we conduct an experiment to explore the potential of WordNet for document clustering and to resolve the conflict introduced by previous works about the values of semantics obtained from WordNet. We tested different similarity measures (e.g. WUP, LCH, RESK), different datasets (e.g. Reuters vs. OHSUMED vs JOURNALS) and different experimental settings (no semantics, with synonyms, with similarity measures).

Results have shown that the clustering results vary depending on the used dataset. If the dataset is heterogenous, comprising of documents related to different domains, the incorporated semantics will not add significant improvement to the clustering results. This is due to the limited coverage of WordNet and inability to measure relatedness between terms that belong to different domains. This result was obvious when using the Reuters dataset.

However, using domain-specific datasets resulted in better clustering results as in the case of OHSUMED and JOURNALS datasets. This indicates that it is easier for WordNet to determine relations between terms that belong to the same domain than to determine between terms related to different domains.

It was also proven that identifying and replacing synonyms produced the best results. The use of similarity measures did not often produce the best results and might sometimes hinder the results. We expect that the similarity scores cause some noise that affect the document's representation. This result conforms with other efforts which also

indicated that, t little or no improvement resulted from using the similarity measures.

In the second part, we proposed an approach to enhance document clustering by leveraging the link structure of Wikipedia. Wikipedia has been chosen because of its hirarichalhierarchal structure, similar to an ontology, and its huge coverage compared to WordNet.

In this work, we proposed an approach for leveraging Wikipedia link structure to improve text clustering performance. Our approach uses a phrase extraction technique to efficiently map document terms to Wikipedia concepts. Afterwards, the semantic similarity between Wikipedia terms is measured by using a measure that is based on the Normalized Google Distance and the Wikipedia's link structure. The document representation is then adjusted so that each term is assigned an additional weight based on its similarity to other terms in the document. Our approach differs from similar efforts from the state of the art in two aspects: first, unlink other works that built their own methods of measuring similarity through the Wikipedia's category links and redirects, instead we used a similarity measure that is modeled after the Normalized Google Distance which is a well-known and low-cost method of measuring similarity between terms. Second, while other approaches used to match all possible n-grams to Wikipedia concepts, our approach is more time efficient as it applies an algorithm for phrase extraction from documents prior to matching terms with Wikipedia. In addition, our approach does not require any access to the Wikipedia's textual content, and relies only on the Wikipedia's link structure to compute similarity between terms. The proposed approach was evaluated by being compared with two different methods (e.g. Bag of Words with no semantics as well

as clustering with WordNet) on two datasets: Reuters 21578 and OHSUMed.

In future work, is recommended to focus on enhancing the clustering of Arabic document by using background knowledge expressed in Arabic. For instance, we may explore the use of Arabic WordNet [e.g. [78]] and the Arabic Wikipedia to measure similarity between Arabic Words.

Regarding the Wikipedia-based clustering approach, a good further aim is to improve the concept-mapping technique: Currently, only the document terms that exactly match Wikipedia concepts are extracted and used for the similarity measure. Instead of exact matching, we aim to utilize the graph of Wikipedia links to build the connection between Wikipedia concepts and the document content even if they cannot exactly match. This approach can be more useful when Wikipedia concepts cannot fully cover the document content.

References

1. Hung Chim, F.X.D.F., *Efficient Phrase-Based Document Similarity for Clustering.* IEEE Trans. Knowl. Data Eng. IEEE Transactions on Knowledge and Data Engineering. 20(9): p. 1217-1229.
2. Rafi, M.S.M.S.F.A., *Document Clustering based on Topic Maps.* IJCA International Journal of Computer Applications, 2010. 12(1): p. 32-36.
3. Sathiyakumari, K., et al., *A Survey on Various Approaches in Document Clustering.* Int. J. Comp. Tech. Appl., IJCTA, 2011. 2(5): p. 1534-1539.
4. D.venkatesan, G.b.a., *Study of Ontology or Thesaurus Based Document Clustering and Information Retrieval.* Journal of Theoretical and Applied Information Technology. 40(1): p. 55-61.
5. Oikonomakou, N. and M. Vazirgiannis, *A Review of Web Document Clustering Approaches*, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Editors. 2010, Springer US. p. 931-948.
6. Yu, L., *A Developer's Guide to the Semantic Web*. 2011: Springer.
7. Patwardhan, S., S. Banerjee, and T. Pedersen, *Using measures of semantic relatedness for word sense disambiguation*, in *Computational linguistics and intelligent text processing*. 2003, Springer. p. 241-257.
8. Srihari, R.K., Z. Zhang, and A. Rao, *Intelligent indexing and semantic retrieval of multimodal documents.* Information Retrieval, 2000. 2(2-3): p. 245-275.
9. Stevenson, M. and M.A. Greenwood, *A semantic approach to IE pattern induction*, in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*2005, Association for Computational Linguistics: Ann Arbor, Michigan, USA. p. 379-386.
10. Meng, L., R. Huang, and J. Gu, *A Review of Semantic Similarity Measures in WordNet.* International Journal of Hybrid Information Technology, 2013. 6(1).
11. Liu, G., et al., *A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge*, in *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011)*2011: Miami Beach, USA. p. 175-178.
12. Recupero, D.R., *A new unsupervised method for document clustering by using WordNet lexical and conceptual relations.* Information Retrieval, 2007. 10(6): p. 563-579.
13. Hotho, A., Staab, S. and Stumme, G. *Wordnet improves text document clustering*. in *in Proc. of the Semantic Web Workshop at 26th Annual International ACM SIGIR Conference*. 2003. Toronto, Canada.
14. Wang, Y. and J. Hodges. *Document clustering with semantic analysis*. in *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*. 2006. IEEE.
15. Jing, L., et al., *Ontology-based distance measure for text clustering*, in *Proceedings of the Text Mining Workshop, SIAM International Conference on Data Mining*2006.
16. Sedding, J. and D. Kazakov. *WordNet-based text document clustering*. in *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*. 2004. Association for Computational Linguistics.
17. Fodeh, S., B. Punch, and P.-N. Tan, *On ontology-driven document clustering using core semantic features.* Knowledge and information systems, 2011. 28(2): p. 395-421.

18. Pantel, P. and D. Lin, *Document clustering with committees*, in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*2002, ACM. p. 199-206.
19. Zhang, X., et al., *A comparative study of ontology based term similarity measures on PubMed document clustering*, in *Advances in Databases: Concepts, Systems and Applications*. 2007, Springer. p. 115-126.
20. Lin, D. *An information-theoretic definition of similarity*. in *ICML*. 1998.
21. Wu, Z. and M. Palmer. *Verbs semantics and lexical selection*. in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 1994. Association for Computational Linguistics.
22. Tversky, A. and I. Gati, *Similarity, separability, and the triangle inequality.* Psychological review, 1982. 89(2): p. 123.
23. Gupta., S.S.a.V., *Recent Developments in Text Clustering Techniques.* International Journal of Computer Applications (0975 – 8887). 37(6): p. 14-19.
24. Han, J., M. Kamber, and J. Pei, *Data mining: concepts and techniques*. 2006: Morgan kaufmann.
25. Li, Y., S.M. Chung, and J.D. Holt, *Text document clustering based on frequent word meaning sequences.* Data & Knowledge Engineering, 2008. 64(1): p. 381-404.
26. Fung, B.C., K. Wang, and M. Ester, *Hierarchical document clustering using frequent itemsets*, in *Proceedings of SIAM international conference on data mining*2003. p. 59-70.
27. Li, Y., C. Luo, and S.M. Chung, *Text clustering with feature selection by using statistical data.* Knowledge and Data Engineering, IEEE Transactions on, 2008. 20(5): p. 641-652.
28. Hammouda, K.M. and M.S. Kamel, *Efficient phrase-based document indexing for web document clustering.* Knowledge and Data Engineering, IEEE Transactions on, 2004. 16(10): p. 1279-1296.
29. Vadivu Ganesan, R.S.a.M.T., *Similarity Measure Based On Edge Counting Using Ontology.* International Journal of Engineering Research and Development. 3(3): p. 40-44.
30. Manjula Shenoy.K, D.K.C.S., Dr. U.Dinesh Acharya, *A New Similarity Measure for Taxonomy Based on Edge Counting.* International Journal of Web & Semantic Technology (IJWesT), 2012. 3(4): p. 23-30.
31. Nguyen, H.A. and H. Al-Mubaid, *A Combination-based semantic similarity measure using multiple information sources*, in *Information Reuse and Integration, 2006 IEEE International Conference on*2006, IEEE. p. 617-621.
32. Petrakis, E.G., et al., *Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies*, in *4th Workshop on Multimedia Semantics (WMS'06)*2006. p. 44-52.
33. Zhang, X., et al., *Medical document clustering using ontology-based term similarity measures.* International Journal of Data Warehousing and Mining (IJDWM), 2008. 4(1): p. 62-73.
34. [cited 2014 1/12/2014]; Available from: http://search.cpan.org/dist/WordNet-Similarity/lib/WordNet/Similarity/.
35. Hirst, G. and D. St-Onge, *Lexical chains as representations of context for the detection and correction of malapropisms.* WordNet: An electronic lexical database, 1998. 305: p. 305-332.
36. Hung, C., S. Wermter, and P. Smith, *Hybrid neural document clustering using guided self-organization and wordnet.* Intelligent Systems, IEEE, 2004. 19(2): p. 68-77.

37. Zhu, S., J. Zeng, and H. Mamitsuka, *Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity.* Bioinformatics, 2009. 25(15): p. 1944-1951.
38. jonathanzong. February 2, 2013 at 1:06am 1/12/2014]; Available from: http://jonathanzong.com/blog/2013/02/02/k-means-clustering-with-tfidf-weights.
39. Yoo, I., X. Hu, and I.-Y. Song, *Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering*, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*2006, ACM. p. 791-796.
40. Banerjee, S. and T. Pedersen. *Extended gloss overlaps as a measure of semantic relatedness*. in *IJCAI*. 2003.
41. Budanitsky, A. and G. Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness.* Computational Linguistics, 2006. 32(1): p. 13-47.
42. Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification.* WordNet: An electronic lexical database, 1998. 49(2): p. 265-283.
43. Pirró, G., *A semantic similarity metric combining features and intrinsic information content.* Data & Knowledge Engineering, 2009. 68(11): p. 1289-1308.
44. Richardson, R., A. Smeaton, and J. Murphy, *Using WordNet as a knowledge base for measuring semantic similarity between words*, 1994, Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University.
45. Yang, D. and D.M. Powers. *Measuring semantic similarity in the taxonomy of WordNet*. in *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*. 2005. Australian Computer Society, Inc.
46. Fodeh, S.J., W.F. Punch, and P.-N. Tan. *Combining statistics and semantics via ensemble model for document clustering*. in *Proceedings of the 2009 ACM symposium on Applied Computing*. 2009. ACM.
47. Moravec, P., M. Kolovrat, and V. Snasel. *LSI vs. Wordnet Ontology in Dimension Reduction for Information Retrieval*. in *Dateso*. 2004.
48. Termier, A., M. Sebag, and M.-C. Rousset. *Combining Statistics and Semantics for Word and Document Clustering*. in *Workshop on Ontology Learning*. 2001. Citeseer.
49. Passos, A. and J. Wainer. *Wordnet-based metrics do not seem to help document clustering*. in *Proc. of the of the II Workshop on Web and Text Intelligente, São Carlos, Brazil*. 2009.
50. Larsen, B. and C. Aone, *Fast and effective text mining using linear-time document clustering*, in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*1999, ACM. p. 16-22.
51. Zhao, Y. and G. Karypis, *Criterion functions for document clustering: Experiments and analysis.* Machine Learning, 2001.
52. Lewis, D.D., *Reuters-21578 text categorization test collection, distribution 1.0.* http://www. research. att. com/~ lewis/reuters21578. html, 1997.
53. Hersh, W., et al. *OHSUMED: An interactive retrieval evaluation and new large test collection for research*. in *SIGIR'94*. 1994. Springer.
54. Group, X.U. 07/15/2005 [cited 2014 1/12/2014]; Available from: http://davis.wpi.edu/xmdv/datasets/ohsumed.html.
55. Salton, G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of*. 1989: Addison-Wesley.

56. **Hotho, A., S. Staab, and G. Stumme, *Text clustering based on background knowledge.* Institute of Applied Informatics and Formal Description Methods AIFB, University of Karlsruche, 2003.**
57. **Group, T.S.N.L.P. 1/12/2014]; Available from: http://nlp.stanford.edu/software/tagger.shtml.**
58. **Varelas, G., et al., *Semantic similarity methods in wordNet and their application to information retrieval on the web*, in *Proceedings of the 7th annual ACM international workshop on Web information and data management*2005, ACM: NewYork. p. 10-16.**
59. **Hotho, A., A. Maedche, and S. Staab, *Ontology-based text document clustering.* KI, 2002. 16(4): p. 48-54.**
60. **Tan, A.-H. *Text mining: The state of the art and the challenges*. in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. 1999.**
61. **Hotho, A., S. Staab, and G. Stumme, *Ontologies improve text document clustering*, in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*2003, IEEE. p. 541-544.**
62. **Huang, A. *Similarity measures for text document clustering*. in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008.**
63. **Wang, P., et al. *Improving text classification by using encyclopedia knowledge*. in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. 2007. IEEE.**
64. **Hu, X., et al. *Exploiting Wikipedia as external knowledge for document clustering*. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009. ACM.**
65. **Huang, A., et al., *Clustering documents using a Wikipedia-based concept representation*, in *Advances in Knowledge Discovery and Data Mining*. 2009, Springer. p. 628-636.**
66. **Cilibrasi, R.L. and P.M. Vitanyi, *The google similarity distance.* Knowledge and Data Engineering, IEEE Transactions on, 2007. 19(3): p. 370-383.**
67. **Hu, X., et al. *Exploiting internal and external semantics for the clustering of short texts using world knowledge*. in *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009. ACM.**
68. **Wang, P. and C. Domeniconi. *Building semantic kernels for text classification using wikipedia*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.**
69. **Mller, C. and I. Gurevych, *Using wikipedia and wiktionary in domain-specific information retrieval*, in *Evaluating Systems for Multilingual and Multimodal Information Access*. 2009, Springer. p. 219-226.**
70. **Gabrilovich, E. and S. Markovitch. *Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge*. in *AAAI*. 2006.**
71. **Spanakis, G., G. Siolas, and A. Stafylopatis, *Exploiting Wikipedia knowledge for conceptual hierarchical clustering of documents.* The Computer Journal. 55(3): p. 299-312.**
72. **Milne, D. and I.H. Witten. *Learning to link with wikipedia*. in *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008. ACM.**
73. **Hu, J., et al. *Enhancing text clustering by leveraging Wikipedia semantics*. in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008. ACM.**

74. **Phan, X.-H., L.-M. Nguyen, and S. Horiguchi.** ***Learning to classify short and sparse text & web with hidden topics from large-scale data collections*****. in** ***Proceedings of the 17th international conference on World Wide Web*****. 2008. ACM.**
75. **Agrawal, R. and R. Srikant.** ***Fast algorithms for mining association rules*****. in** ***Proc. 20th int. conf. very large data bases, VLDB*****. 1994.**
76. **Agrawal, R., T. Imieliński, and A. Swami.** ***Mining association rules between sets of items in large databases*****. in** ***ACM SIGMOD Record*****. 1993. ACM.**
77. **Miner, W.** ***Wikipedia Miner*****. 20th October 2014]; Available from: http://wikipedia-miner.cms.waikato.ac.nz/.**
78. **Elkateb, S., et al.** ***Arabic WordNet and the challenges of Arabic*****. in** ***Proceedings of Arabic NLP/MT Conference, London, UK*****. 2006. Citeseer.**