ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

MEASURING SIMILARITY BETWEEN NEWS ITEMS USING LINK ANALYSIS AND SEMANTIC APPROACH

By: Yemane Seged Gebru

A THESIS SUBMITTED TO

THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN

PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE INCOMPUTER SCIENCE

August, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

MEASURING SIMILARITY BETWEEN NEWS ITEMS USING LINK ANALYSIS AND SEMANTIC APPROACH

By: Yemane Seged Gebru

ADVISOR:

Fekade Getahun (PhD)

APPROVED BY

Examining Board:

1. Dr. Fekade Getahun, Advisor _____

2. Dr. Mulugeta Libse, Examiner _____

3._____

I would like to dedicate my thesis to my **family** and my **friends**.

ACKNOWLEDGMENT

I have received a lot of kind support from many loved ones and friends when I worked on my thesis. I would like to take this opportunity to pass my regards.

First and foremost I thank God and His Mom for they are always besides me during all hardships. I am also grateful to my home university, Mekelle University, for this golden opportunity.

I am indebted to my advisor **Dr. Fekade Getahun** for his all rounded support, his patience, careful supervision, and tireless encouragement at all times.

I would like to thank my friend Ephrem Berhe, who has been working with and helping me to study and understand C# programming language. I would also like to thank Ato Zelalem Assefa and his colleagues from the MOE Tele-education data center for his genuine support in providing me with the necessary facilities. Also, my heartfelt regards go to my friends Ashenafi Gebre, Aberham Woldu, Luel Berhe, Goitom Tegeng and all others and my beloved friends. They have been of great help, support, and encouragement in accomplishing this research.

Above all, I am deeply thankful to my Father Seged Gebru, My mother Mulu Medhaniye, and My sisters Berekti, Tirhas and Berhan, who supported me in each and every day. Without their everlasting love, support and encouragement, this thesis would have never been completed

.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

C#	(C-Sharp)
DTD	Document Type Definition
HTML	Hyper Text Markup Language
IC	Information content
IL	Intermediate language
IR	Information Retrieval
KB	Knowledge Base
LCS	Lowest Common Subsumer
NLP	Natural Language Processing
POS	Part Of Speech
P-Rank	penetrating Rank
RSS	Really Simple Syndication
TED	Tree Edit Distance
TF-IDF	Term Frequency-Inverse Document Frequency
TR	Text relatedness
URL	Uniform Resource Locator
www	world wide web
w3c	www consortium
WG	Working Group
XML	Extensible Markup Language

ABSTRACT

In the recent years, the ways people acquire information have been completely changed. Activities such as reading hardcopy materials such as books, journals, and newspapers, have radically declined, and most of the people go online to find recent and up-to-date information. As a result, news feeds technology such as RSS and ATOM was created to allow news users to get frequently update information. However, the number of news items that will be downloaded to the aggregator will be unmanageable when the number of provides grows. This will be even annoying when some of the news items are similar to already read news items.

One of the possible solutions to this challenge is to measure similarity among news items. Measure similarity between news items is pre-requisite to a number of application areas, grouping, clustering, merging and revision/version control. Since news Feeds are XML files, they do have several sub-elements such as title, description/summery, link, guild, etc…. Previously item/entry sub-elements such as title and description/summary have been used as input in measuring similarity. In this work, we propose to use link sub-element information that improves and supplement the similarity computation between two items. As news page contains links to set of related news pages, our new similarity approach uses these links in measuring similarity. We developed new similarity measures that consider the link sub-element and related news links together with their anchor text.

In order to validate our approach, we developed a prototype implementing the link based news Feed similarity measure. Experimental results show that the link based news feed similarity is more helpful in measuring similarity when it is combined with computing similarity only with title and description sub-elements and compared to using SimRank and co-citation.

Keywords: similarity measure, link analysis, news Feed, Semantic similarity

CHAPTER ONE

INTRODUCTION

1.1 Introduction

This chapter is going to discusses background to the problem area in Section 1.1, motivation for the work in Section 1.2, statement of the problem in Section 1.3, general and specific objectives of the study in Section 1.4, research methodology in Section 1.5, scope and limitation of the study in Section 1.6, significance of the study in Section 1.7 and finally application results of the study are discussed in Section 1.8.

1.2 Background

In the recent years, the ways people acquire information have been completely changed. Activities such as reading hard copy materials such as books, journals, and newspapers, have radically declined, and most of the people go online to find recent and up-to-date information.

News feeds are Extensible Markup Language (XML)-based technologies that makes easy to manage web data flow and allow web users to get frequently updated information [**1**], [**2**], [**3**]. Currently, RSS (Really Simple Syndication) and ATOM are the two popular syndication formats [**4**].Both formats contain some descriptive information extracted from a fully prepared article such as title, date of publication, a link to the original article, guild which is a unique identifier to the feed and a summary or a description text.

The use of RSS/news feed aggregators empowered web users to get information at click away rather than roaming from site to site. (i.e., the user registers the address of their favorite content provider in the aggregator for instance Google Reader and the aggregator download the recently published feed as soon as it gets).However, the number of news items that will be downloaded to the aggregator will be unmanageable when the number of provides grows. This will be even annoying when some of the news items are similar to already read news items.

One of the possible solutions to this challenge is to measure similarity among news items. Measure similarity between news items is pre-requisite to a number of application areas, grouping, clustering, merging and revision/version control [**5**].

In measuring similarity, researchers [**6**], [**7**], [**5**]proposed to use text base descriptors of Item/entry such as title and description/summary as input main sources for measuring similarity. However, there is important information embedded in the link element that improves and supplement the similarity computation and helps to infer similarity between the items. The aim of this thesis is to use the link sub-element as an additional component in measuring the similarity between News items.

1.3 **Motivation**

News feed's link element contains URL (Uniform Resource Locater) that contains the address of the actual news source. URL contains valuable text information such as hierarchal folder names representing path and file name besides the domain name. In addition to this, most News pages contain pointers to related news pages that enrich the content of the current News page. Each of these related links also contain anchor text summary of the page it is pointing to.

We take into consideration the text information embedded in link element of the News feed and its recommended related hyperlinks in order to get a better similarity score.

To motivate our work, let us consider CNN1 and BBC1 news feeds extracted from CNN and BBC and shown in Table 1-1.

Table 1-1 News feeds extracted from CNN and BBC

CNN1	<item><title>Alarm over Sudan, South Sudan clashes</title><description>U.S. and international powers warn that fighting between Sudan and South Sudan could lead to war, less than a year after South Sudan became independent.</description> <link> http://edition.cnn.com/2012/03/28/world/africa/sudan-violence/</link></item>
BBC1	<item><title>Somali piracy: EU forces in first mainland raid</title><description>EU naval forces have conducted their first raid on pirate bases on the Somali mainland, saying they have destroyed several boats.</description><link>http://www.bbc.co.uk/news/world-africa-18069685</link></item>

Considering item descriptor[1] defined on title and description element, the similarity between CNN1 and BBC1 while considering semantic information is close to zero. However, as both news items are describing event occurring in Africa they possess some degree of similarity caused by URL text like World, Africa and Sudan. This similarity is more visible when the similarity between hyperlinks located in the source page is considered.

1.4 Statement of the Problem

Since, pervious works on news feed similarity approaches do not utilize the link sub-element in the computation process. In this thesis, we assess and provide an approach that measure similarity using link sub-element. And the following are the key research questions associated with the thesis.

- How to identify key contents of link element?
- How to identify hyperlinks of pages related to given news?
- How to measure similarity between news items based on the similarity between its component link elements?

1.5 Objectives of the Study

1.5.1 General Objective

The main objective of this study is to provide an approach that measures the similarity between News items using link sub-element information

1.5.2 Specific Objective

The specific objectives of this work are:

- Understand the different technique in identifying the component of hyperlinks
- Assess the capability and drawback of existing link based similarity methods
- Propose an approach that extend the existing link similarity measures to be semantic-aware

[1] Item Descriptor [**5**] is set of element names used in computing RSS news item similarity.

- Evaluate the capability of our measure with prototype
- Compare the capability of the measure against existing link-based similarity approaches

1.6 Research Methodology

In order to achieve the specified objectives of the research, the researcher is going to use different methodologies for various stages of this thesis work.

1.6.1 Review of Related Literature

At first, issues and areas related to the thesis are going to be reviewed. This is done mainly through reading journal papers, articles, books and other reading materials that enrich the understanding of the subject area. Major activities performed in this phase were:

- Reviewing major works in news feed similarity measure, news merging and the various issues and challenges that are raised along with it. Since the area is relatively new, it needed to be studied in greater depth so as to address the various areas of research currently underway.
- Review major works in the area of link similarity.
- The main concern of this thesis work is to use link sub-element of a given news item/entry and its outgoing recommended link textual information in measuring similarity between two news items. Hence, a comprehensive study of link text extraction, outgoing link extraction, their nature and characteristics and other related issues are going to be performed.

1.6.2 Develop Algorithm

Secondly, we are going to develop a new algorithm for the new proposed solution. The main input for our algorithm development is going to be the different reviewed literatures. Discussion with advisors and with colleagues who work in related areas is also going to be the other important input. Here the major activities is going to be clearly specifying the input, process and output values for the algorithm to be developed.

1.6.3 Develop Prototype

At last we are going to deal with the design and development of a prototype that measures similarity between two news items/entry based on their link sub-element. To successfully develop our prototype, we will use different programming language, database management system software, etc.

1.7 Scope

Even though the findings of the research is important for other different areas of web applications, the scope of this research is limited to measuring similarity between two news items using their link sub-element and their related links extracted from news page.

1.8 Significance of the Study

These days news feed technology are becoming one of the frequently accessible information throughout the internet. Determining similarity of two news feed will be used as input to handle issues related to news feed analysis. As a result, the output of research will have impact in avoiding redundant news, improving clustering quality, creating relevant grouping, merging news feed and it is also going to be an input to other applications.

1.9 Application of the Results

The result of this thesis could be used to support the existing similarity measure and can be considered as an additional alternative similarity measure mechanism of the previous works [**5**]. Our approach takes the advantage of link sub-element information from News item/entry (internal) and it also considers outgoing related URL anchor text information (external) embedded in the source news supplemented with knowledge base that contains collection of related concepts. These uses help us to infer the related news between news items.

The results from this study could add additional concept set to the news feed similarity from the external part of the news feed i.e. related link information at different depth of the web graph. This means that it increases the concept set of the similarity measure external link information.

1.10 **Thesis Organization**

The remaining part of this document is organized in 5chapters. Chapter 2 lays the foundation for other parts of this research. It describes the necessary background for the rest of the work. It thoroughly discusses important literatures in the area of similarity definition and link similarity upon which this thesis work builds. Chapter 3looks into different kinds of efforts that have been done in the area of text based similarity measure.

Chapter 4 details our similarity/relatedness measure between a pair of News item based on their Anchor text and URL text. Chapter 5presents the experimentation phase of the study at hand. Results of the similarity measure experiments were also discussed here. Finally, Chapter 6concludes the thesis report by describing conclusions, contribution and our future research directions.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter presents the review of literatures upon which this thesis work is basing on. We present major works and issues related to similarity. Overview of measuring similarity and relatedness is discussed in Section 2.1 and Link-Based similarity measures in Section 2.2upon which this thesis work builds.

2.2 Measuring Similarity

Before reviewing different literatures and describing previous efforts conducted by different researchers, it is important to define what do we mean by similarity and Relatedness? Based on the WorldNet Semantic English dictionary [**8**], similarity means: "the quality of being similar". And relatedness means: “A particular manner of connectedness”.

Measuring similarity between things is a fundamental and widely used concept. It has been a subject of great interest in human history since a long time ago. Even before computers were made, humans have been interested in finding similarity between objects and as a result many different similarity measures have been proposed in different disciplines such as mathematics[**9**] psychology [**10**].

By using the computer it has been easier to what extent two or more objects are similar to each other. In order to specify what kind of similarity and for what purpose is needed, it can be separated into different categories.

Computing similarity plays a key role in different computer research areas such as Natural Language Processing (NLP), Information Retrieval (IR), Information Integration and Machine Learning.

In [49], the authors presented an information theoretic definition of similarity and they demonstrated also how their definition can be used to measure the similarity in a number of different domains by clarifying their perceptions about similarity as follows:

- The similarity between A and B is related to their **commonness**. The more commonness they share, the more similar they are.
- The similarity between A and B is **related** to the differences between them. The more differences they have, the less similar they are.
- The maximum similarity between A and B is reached when A and B are **identical**, no matter how much commonness they share.

So similarity is the ratio between the amount of information in the commonality and the amount of information in the description of A and B. If commonality of A and B is known, their similarity would tell how much more information is needed to determine what A and B are [49].

In the next section we present the litrature review of link-based similairty measures.

2.3 Link-based Similarity Measures

In recent years a considerable amount of research works [**11**] , [**12**] and [**13**]have focused on examining collections of hyper-linked pages and structures called link analysis. Link analysis[2]techniques have been used in extracting knowledge, by measuring similarity between objects (for example web pages and citation among articles).Currently, analyzing web link structures is being widely used in many applications for the purpose of obtaining similar objects. For example, different social network applications such as Facebook, Google Plus and Twitter use link analysis in order to suggest relevant objects in their respective social network[**14**], [**15**].

The different link-based similarity approaches are categorized into two groups neighbor-based and graph-based.

2.3.1 Neighbor-based Link Similarity Measure

The intuition behind neighbor[3]-based methods is “similar objects have similar neighbors.” They focus on comparing local neighborhood structures of the given objects. Traditional methods

[2]Link analysis is a network analysis that explores key relationships among objects.

[3] A neighbor of link in web graph is any parent link (incoming link) or child (outgoing) link.

contain Co-citation [**16**], Bibliographic coupling [**17**], Jaccard Measure [**18**], SimRank [**19**], and rvc-SimRank [**13**].

Co-citation

Initially Co-citation was proposed to measure the similarity between scientific papers [**16**]. The similarity between two articles is dependent on articles that cited both of them. This reflects the assumption that the author of a scientific paper cites only papers related to her/his work. Thus, if paper X and Y are both cited by paper Z, then both are related in some sense to one another, even if they do not directly cite each other. On the other hand if paper X and Y are cited together in many papers, it means that X and Y have a strong relationship or similarity. Later this assumption of co-citation is applied to compute the similarity between web documents considering links as citations (i.e., "a Web page author will insert links to pages related to her/his own page").

For example, given two web pages x and y their corresponding similarity is dependent on the number of common links (i.e., links referenced in both x and y) over the total number for links. It is formalized as follows:

$$\mathrm{Sim}_{\mathrm{CoCite}}(X,Y) = \frac{|I_X \cap I_Y|}{|I_X \cup I_Y|} \quad (1)$$

Where

- I_x, I_y links to web page X and Y (i.e., ingoing links) respectively.
- $|I_X \cap I_Y|$- the number of shared ingoing links shared by X and Y
- $|I_X \cap I_Y|$ – total number of ingoing links

The $\mathrm{Sim}_{\mathrm{CoCite}}(X,Y) \in [0,1]$ and the higher the number of shared links the higher the similar value.

One basic limitation of co-citation is that it computes similarity between links only when they are commonly referenced by a link. In other words it ignores outgoing links.

Bibliographic Coupling

In [**17**], the authors presented a complementary similarity measure approach that uses bibliographic information called bibliographic coupling similarity measure. Their measure is based on the fact that the similarity between papers is dependent on the number of resource both cited i.e., "two documents share one unit of bibliographic coupling if both cite the same paper". This principle can be applied to measure the similarity between web pages. That is two web pages

focusing on related domain are likely to refer to the same pages. Thus, two web pages have a bibliographic coupling if both refer the same page. More formally, given two web pages x and y and O_x and O_y representing the set of links extracted from x and y (O_x and O_y are also called outgoing links) respectively, the similarity is defined as the ratio of the number of common links over the total number of links referenced by X and Y. It is formalized as follows: d as:

$$\mathrm{Sim}_{\mathrm{Biblio}}(X, Y) = \frac{|O_X \cap O_Y|}{|O_X \cup O_Y|} \quad (2)$$

Where:

- $|O_X \cap O_Y|$ – the number of common referenced links of X and Y
- $|O_X \cup O_Y|$- the total number of links referenced by X and Y

According to the Equation, the more common children pages X and Y have, the more related they are. This value is normalized by the total set of children, to fit between 0 and 1. If both and are empty, the bibliographic coupling similarity will be zero.

Bibliographic coupling also has limitation. It computes similarity between links only when they are commonly referring a link. In other words it ignores outgoing links. Latter on Amsler [**20**] is proposed and discussed in the following section.

Combination of Citation and Bibliographic Coupling (Amsler)

In order to take advantage of the information available in citations in measuring similarity between articles [**20**] Amsler proposed an approach that combines both co-citation and bibliographic coupling. Thus two papers X and Y are related if

- X and Y are cited by the same paper,
- X and Y cite the same paper, or
- X cites a third paper Z that cites Y.

As for the previous measures, the Amsler similarity measure can also be applied to measuring similarity between Web pages, through replacing citations by links. Given two web pages X and Y, I_x and I_y be the set of parents (in-links) of X and Y respectively, and O_x and O_y be the set of children (out-links) of X and Y respectively, the Amsler similarity X and Y is defined as:

$$\mathrm{Sim}_{\mathrm{Amsler}}(X, Y) = \frac{|(I_x \cup O_x) \cap (I_y \cup O_y)|}{|(I_x \cup O_x) \cup (I_y \cup O_y)|} \quad (3)$$

The equation tells us that, the more links (either in-link or out-link) X and Y have in common, the more they are related. The measure is normalized by the total number of links. If neither X nor Y have any in-link or out-link, the similarity is defined as zero.

Even though, Amsler similarity measure is able to solve the limitation of co-citation and bibliographic coupling by considering both in-links and out-links, it ignore the similarity between neighbors. This is going to reduce its performance [**21**].

Jaccard Link Similarity Measure

Jaccard link similarity measure also known as the Jaccard's Coefficient. It is defined as the size of the intersection (this include all neighbors in-link and out-link) divided by the size of the union of two sets [**18**]. For a given two objects X and Y, let (x) and (y)denote their respective neighbor sets, therefore the similarity is defined by:

$$\mathrm{Sim}_{\mathrm{Jaccard}}(X, Y) = \frac{|(x) \cap (y)|}{|(x) \cup (y)|} \quad (4)$$

Even though, Jaccard link similarity measure is able to solve the limitation of co-citation and bibliographic coupling by considering both in-links and out-links, it ignore the similarity between neighbors. This is going to reduce its performance [**21**].

SimRank

SimRank is one of the general link similarity measures. It can be applicable in any domain with object-to-object relationships, that measures similarity of the structural context in which objects occur, based on their relationships with other objects [**19**]. The intuition behind SimRank is "two objects are similar if they are related to similar objects." It improves the accuracy of Co-citation, in which the similarity score between two web pages is defined by the number of in-link neighbors that they have in common. SimRank computes similarity iteratively.

SimRank Equation: Given two objects a and b, their similarity is denote by$\mathrm{Sim}_{\mathrm{Rank}}(a, b) \in [0,1]$ and defined as:

$$\mathrm{Sim}_{\mathrm{Rank}}(a,b) = \frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}\mathrm{Sim}_{\mathrm{Rank}}(I_i(a), I_j(b)) \quad (5)$$

Where:

- $I(a)$ denotes the set of in-link pages of a.
- C is a constant between 0 and 1
- $|I(a)||I(b)|$is the number of all possible neighbor pairs.

Notice that if I(a) = ∅ or I(b) = ∅ then $\mathrm{Sim}_{\mathrm{Rank}}(a,b) = 0$, otherwise the result will b the sum of similarity between overall neighbors pairs $I_i(a)$ and$I_j(b)$ withich is $\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}\mathrm{Sim}_{\mathrm{Rank}}(I_i(a), I_j(b))$.

Therefore $\mathrm{Sim}_{\mathrm{Rank}}(a,b)$ is the product of constant C times the average similarity over all possible neighbor pairs between $I(a)$ and $I(b)$.

RVC-SimRank

Usually without scanning the entire web Graph[4], it is not easy to know all in-links referencing a given web document. However, a web page has a good knowledge of out-links (i.e., address of pages/ resources) it is referencing. Based on this fact, the authors in [**13**] proposed a reverse-SimRank (also called RVC - SimRank) that improves Bibliographic coupling.

Given two objects a and b, their similarity is denote as$\mathrm{Sim}_{\mathrm{RVC-Rank}}(a,b) \in [0,1]$. A recursive equation for $\mathrm{Sim}_{\mathrm{RVC-Rank}}(a,b)$ is provided in Equation below. i.e. if a = b then $\mathrm{Sim}_{\mathrm{RVC-SimRank}}(a,b)$ is defined to be 1. Otherwise,

$$\mathrm{Sim}_{\mathrm{RVC-SImRank}}(a,b) = \frac{C}{|O(a)||O(b)|}\sum_{i=1}^{|O(a)|}\sum_{j=1}^{|O(b)|}\mathrm{Sim}_{\mathrm{RVC-SimRank}}(O_i(a), O_j(b)) \quad (6)$$

Where:

- $O(a)$denotes the set of out-link pages of a.
- C is a constant between 0 and 1
- $|O(a)|, |O(b)|$is the number of all possible out-going neighbor pairs.

[4]Web graph is a directed graph that represents the Web.

Notice that if O(a) = ∅ or O(b) = ∅ then $\mathrm{Sim}_{\mathrm{RVC-SimRank}}(a, b) = 0$otherwise the result will be the sum of similarity between overall neighbors pairs$O_i(a)$ and$O_j(b)$.

Therefore $\mathrm{Sim}_{\mathrm{RVC-Rank}}(a, b)$ is the product of constant C times the average similarity over all possible neighbor pairs between $O(a)$ and $O(b)$.

All the classical neighborhood similarity measure are very efficient and easy to implement, and are being used in different applications. But for vast amount of data sources like the Web, considering only direct neighbors is obviously not enough. Alternatively, SimRank makes an extension by taking similarity between neighbors into account. However, it has some challenge [**22**], which may influence the output score.

Latter on PageSim [**13**], MatchSim [**21**] , P-Rank (Penetrating Rank) [**12**], and C-Rank, [**23**] are proposed and overcome the drawbacks of traditional neighbor-based methods in different ways.

PageSim

PageSim [**13**] is one of the known link-based similarity measures that extends co-citation algorithm. The similarity score between two web pages is defined as the number of their shared/common in-link neighbors using the PageRank score propagation principle.

PageRank [**24**] is one of the most ranking algorithm which assigns global ranking scores to all web pages. The authors of [**13**] takes PageRank's score of a web page as the importance (weight or similarity score) of it in the PageSim method. PageSim approach is described as follows:

- First, each web page only contains its own similarity score, and then propagates its own similarity score to it's out-link neighbors.
- After the propagation, each page will have its own similarity score as well as the similarity scores of others.
- Finally calculate the PageSim score of each pair of pages by "summing their common similarity scores up".

PageSim equation:

- Given a directed graph $G = (V, E)$ with vertices V representing web pages vi ($i = 1,2,3 \ldots n = |E|$) and directed adge E which represent hyperlink between web pages
- Let PR(v) denotes the PageRank score of page v, for $v \in V$

- Let $PG(u,v)$ denotes the PageRank score that page u propagates to page v through PATH(u, v)

$$PG(u,v) = \sum_{p \in PATH(u;v)} \frac{PR(u)}{\prod w \in p; w \neq v \, |O(w)|} \quad (7)$$

The PageSim between u and v denoted as PS (u, v) and formalized as:

$$PS(u,v) = \sum_{i=1}^{n} \min(PG(v_i,u), PG(v_i,v)) \quad (8)$$

Where: u, v ∈ V

P-Rank (Penetrating Rank)

P-Rank [**12**]improves Amsler measure between pair of objects as a weighted sum of the similarity scores computed using rvs-SimRank and SimRank. The authors showed that P-Rank is semantically complete and includes all the well-known similarity measures, including Co-Citation, Bibliographic Coupling, Amsler and SimRank.

More specifically, the two main assumptions of P-Rank are:

- Two entities are similar if they are referenced by similar entities
- Two entities are similar if they reference similar entities.

P-Rank similarity score between a and b denoted as $Sim_{P-Rank}(a,b) \in [0,1]$ and formalized as, if $= bs(a,b) = 1$, and when $a \neq b$ when can use the following equation.

$$Sim(a,b) = \lambda \times \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} Sim\left(I_i(a), I_j(b)\right) + (1-\lambda) \frac{C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} Sim(O_i(a), O_j(b)) \quad (9)$$

Where

- $\lambda \in [0,1]$ is a weight given to the similarity between in-link and (1-λ)is weight for similarity between out-link directions which expresses the relative weight of similarity computation between in-link and out-link directions,
- $C \in [0,1]$is set as damping factor.

By changing the value of C and λ, the P-Rank representsany of the knownlink-based similarity measures. For example

- C = 1, and λ = 1 (or λ = 0), P-Rank represents Co-citation (or Coupling).
- C = 1 and λ = 0.5, P-rank represents Amsler.

MatchSim

MatchSim [**21**] focus on the neighbor-based approach that bases on the assumption that "similar objects have similar neighbors. It is proposed as an extension and enhanced version of the traditional neighbor based link similarity measures.

Given two objects a and b in a graph of size n, construct weighted bipartite Graph[5]→ $G_{a,b} = (I(a), I(b), E, w)$,

Where:

- Edge$E = \{(u,v)|u \in I(a), v \in I(b)\}andw(u,v) = sim(u,v)$.

The MatchSim score is defined by:

$$sim(a,b) = \frac{\widehat{W}(a,b)}{\max(|I(a)|,|I(b)|)} \quad (10)$$

Where:

- $\widehat{W}(a,b)$ denotes the weight of a maximum matching between$I(a)$and$I(b)$and it is computed as

$$\widehat{W}(a,b) = W(m_{ab}^{*}) = \sum_{(u,v)\epsilon m_{ab}^{*}} sim(u,v) \quad (11)$$

Where:

- $m_{ab}^{*}$is a maximum matching between$I(a)$and $I(b)$

[5] A bipartite graph (or bi-graph) is a graph whose vertices can be divided into two disjoint sets U and V such that every edge connects a vertex in U to one in V; that is, U and V are independent sets. Equivalently, a bipartite graph is a graph that does not contain any odd-length cycles

Before calculating m^*_{ab}MatchSim always normalizeI(a) and I(b) to have equal size.This is because "if the graph is not completely bipartite (a and b are not of equal size), fake vertices and zero-weighted edges are inserted to make up the missing part" [**21**]. Therefore, before computing the $m*$, converting a and b to be "equally sized" is necessary, and it is defined as:

$$L_{ab} = |m^*_{ab}| = \max(I(a), I(b)) \quad (12)$$

C-Rank

C-Rank is a similarity measure designed for scientific literature databases. It uses both in-links and out-links disregarding the direction of references. C-Rank[**23**] measures the similarity between two papers p and q following the following three cases:

- p and q are similar if they have high number of out-going links in common;
- p and q are similar if they have high number of in-going links in common and
- p and q are similar if many of the papers that are referenced by p reference q.

Among the three cases, the first and the second cases are captured in Bibliographic Coupling and Co-citation, respectively; however these two measures fail to address both cases simultaneously. Moreover, the authors showed that none of the similarity measures is capable to address the third case.

C-Rank uses both in-links and out-links at the same time. Similar to that the accuracy of Co-citation and bibliographic (Coupling) is improved by iterative SimRank and rvs-SimRank, C-Rank is defined iteratively. Equation 13 represents C-Rank:

$R_k(p, q)$, denotes the similarity score between p and q at iteration k. At iteration zero the similarity score is 0 when p and q are not equal otherwise1.Else if iteration greater than zero

$$R_{k+1}(p, q) = \frac{c}{|L(p)||L(q)|} \sum_{i=1}^{|L(p)|} \sum_{j=1}^{|L(q)|} R_k(L_i(p), L_j(q)) \quad (13)$$

Where,

- L(p) denotes the set of undirected link neighbors of paper p.
- L(q) denotes the set of undirected link neighbors of paper q.

2.3.2 Graph-Link Similarity Measures

Unlike neighborhood link similarity measure graph-based link similarity measures uses the whole global structure of graph into consideration. Most of the measures in this category are used to rank a web pages or a links in the web. Some of the prominent measures in this category are PageRank, HITS algorithm [67], and Companion [31]

PageRank

PageRank is an algorithm used by Google search engine, originally it was formulated by Sergey Brin and Larry Page[**24**]. The principle behind this algorithm is the world of academia (academic world). In the academia, the importance of a research paper is judged by the number of citations the paper has from other research papers. PageRank uses this principle to rank web documents f based on the number of hyperlinks pointing to it (also known as inbound links) from other web pages. Therefore, according to the PageRank concept, the rank of a document is given by the rank of those documents which link to it. The rank of these documents again is given by the rank of documents which link to them.

HITS

HITS stands for Hypertext Induced Topic Search [**25**], and unlike PageRank which is a static ranking algorithm, HITS is search query dependent. When the user issues a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings of the expanded set of pages, authority ranking and hub ranking. An authority is a page with many in-links. The page may have good or authoritative content on some topic. A hub is a page with many out-links. The page serves as an organizer of the information on a particular topic and points to many good authority pages on the topic. The key idea of HITS is that a good hub points to many good authorities and a good authority is pointed by many good hubs. Therefore, Authorities and hubs have a mutual reinforcement relationship.

Companion Algorithm

Companion algorithm is derived from the HITS algorithm proposed by Kleinberg for ranking search engine queries [**26**]. HITS algorithm could be used for finding related pages as well, and provided subjective evidence that it might work well. On a different approach, [**27**]proposed the

Companion algorithm. Given a Web page d, the algorithm finds a set of pages related to d by examining its link structure. Companion is able to return a degree of how related each page is to d. This degree can be used as a similarity measure between d and other pages.

2.3.3 URL and Anchor Text Information

The above link-based similarity measure, they only consider the neighborhood and the whole graph structure of a web. However it is important to look the internal and external textual information of a link.

Web page links contain textual information in their surroundings. Particularly from the HTML the anchor element/tag <a>, we can get useful textual data such as Anchor text and URL text information. On one hand, anchor texts in Web documents provide a short description of the target document. Although they are initially created to help users navigate from one page to another, they usually provide an additional and complementary description of the document contents [**28**]. On the other hand, anchor texts also share similar characteristics with Web queries, for example, they are usually short and descriptive. They have a better chance to match user queries than the content words of a document [**29**].

For instance, the anchor element South Korea, contains a hyperlink to the Reuters News web site, which leads users to visit Full coverage of South Korea. In this example, the anchor contains two types of information:

- anchor text "South Korea";
- URL http://www.reuters.com/places/south-korea

An anchor text provides a good description of the page

At the same time the URL text “places/south-Korea” is also a good description of page "http://www.reuters.com/places/south-korea". After performing some pre-processing text like (extract concepts/works other than those representing domain information and embedded in the URL) stop word removal, change wild cards to blank space, path separators such as“/” and “-”to blank space), we can get a textual descriptor “places south Korea” which is describe semantically what the anchor text “North Korea” is.

Therefore, in this thesis we can use the anchor and URL text of a hyperlink as an important element in computing the similarity between pair of news items.

2.3.4 Recommendation Links Textual Information

In addition to the link sub-element of the News item/entry, there may be hyperlinks in the detailed content of the news, which are recommended by the News provider. Figure 2-1 shows a news feed item published by BBC. This page has a title: “Owen Coyle: FabriceMuamba sent Bolton message of support”, description: “Bolton Wanderers manager Owen Coyle has revealed that his players received a message of support from FabriceMuamba before their crucial win against fellow strugglers Blackburn Rovers at the Reebok” and

Link: “http://www.bbc.co.uk/sport/0/football/17501423”

```
<title>Owen Coyle: FabriceMuamba sent Bolton message of support</title>
<description>Bolton Wanderers manager Owen Coyle has revealed that his players
received a message of support from FabriceMuamba before their crucial win
against fellow strugglers Blackburn Rovers at the Reebok.</description>
<link>http://www.bbc.co.uk/sport/0/football/17501423<link>
```

Figure 2-1 News feed from the Guardian news page

The link element contains the actual news page that contains link to a number of related pages. For instance, Figure 2-2 shows list of links extracted from the news item shown in Figure 1.3 with associated anchor text obtained from the pages of links at level one.

Link	Anchor text
http://www.bbc.co.uk/sport/0/football/17417973	cardiac arrest he suffered during the FA Cup quarter-final against Tottenham
/sport/0/football/17412469	gave Bolton an emotionally-charged 2-1 win
/news/uk-england-manchester-17498947	shown from both sets of supporters
/sport/0/football/17412469	Bolton2 - 1Blackburn
/sport/0/football/premier-league/	Premier League
/news/uk-england-manchester-17498947	Bolton fans stage Muamba tribute
/news/england/manchester/	Manchester
http://www.bbc.co.uk/sport/0/football/17370584	Football on the BBC
http://www.bbc.co.uk/manchester/sport/index.shtml	BBC Manchester sport
http://www.bbc.co.uk/weather/5day.shtml?id=1211	Bolton weather
http://news.bbc.co.uk/sport1/hi/football/eng_prem/default.stm	BBC Sport Premier League
http://www.bwfc.premiumtv.co.uk/page/Welcome -	External siteBolton Wanderers

Figure 2-2Samplelinks and anchor text extracted from sample item at level one

Thus measuring similarity between news items should take into consideration the similarity between links at different levels.

2.4 **Summary**

In this Chapter, we presented two prominent categories of link-based similarity measure-Neighbor and Graph-based similarity measures. Unlike graph-based, neighbor-based similarity measures are classical and have limitation to consider the whole graph structure of a web. Both of these link similarity measures only consider the link structure of the web graph during computing similarity. However, it is so important to investigate and use the textual information embedded internally and externally in the link element. After extracting the textual information of the links the next step is to measure similarity as the similarity between textual values. The next chapter discusses the prominent works related to the thesis work.

CHAPTER THREE
RELATED WORK

3.1 Introduction

RSS is a rich text XML document that describes news items published by news providers. The known approach for measuring the similarity between news items is based on text similarity, XML similarity. These measures will be presented in this Chapter.

The rest of the Chapter is organized as follows. Section 3.1 discuses text pre-processing techniques made before measuring similarity, Section 3.2 reviews the most known text based similarity measures. Section 3.3 assesses XML document similarity measure and finally Section 3.4 summarizes the Chapter.

3.2 Text Pre-processing

Before measuring the similarity between two texts, usually text pre-processing technique is employed. Text pre-processing consists of different tasks including tokenization, stop-word removal, and stemming. In [**30**] the task of preprocessing defined as removing stop words and word stemming.

In the English language, there are terms that appear very frequently on the collection of documents and most of which are not relevant for the information retrieval tasks such as measuring similarity between two texts. These include articles, conjunctions, prepositions, etc (for example, the words “a”, “an”, “are”, “be”, “for” …) are referred as English stop words.

Stemming or lemmatizing is the process of reducing inflected (or sometimes derived) words into their stem. For example, words “reducing” and “reduced” are reduced to their stem word “reduce”. Since all the variants of the word will be the same after being stemmed measuring similarity becomes easier to process and produce more accurate result and save storage space for the program. There are many stemming algorithms among which Porters [**31**].

On the other hand, Tokenization is the process of demarcating and possibly classifying sections of a string of input characters. According to [**32**] token identification task is used to automatically identify the boundaries of a variety of phrases of interest in raw text and mark them up with

associated labels. The initial task is to produce a list of attributes. These attributes could be single words or word phrases.

3.3 Text Similarity Measure

Text similarity is the one among the several kinds of similarity measures between two objects. It is about measuring two or more strings with each other to find out how similar they are. The different text similarity methods can be categorized into two: Classical (Syntactic) text similarity and Semantic-based text similarity measure.

3.4 Classical (Syntactic) Text Similarity Measure

There are a number of classical text similarity measure algorithms. This sub section discusses the most known algorithms: Vector Space Model and Edit Distance.

3.4.1.1 Vector Space Model

In Vector Space Model, texts are preprocessed, the distinct terms/words in the text are represented in vector and every dimension corresponds to a separate term or token. Each term in the vector will have a corresponding weight which is computed using Term Frequency, Inverse Document Frequency (TF-IDF) [**33**]. Once texts are represented the similarity between texts can be computed using any vector based method such as Cosine similarity, Jaccard Similarity Coefficient and Euclidean Distance.

Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF calculates how many times a term occurs in a given text. When we are looking at a document, the number of times a given term occurs in a single document is called the term frequency [**34**].

To understand TF-IDF let us see TF and IDF separately. TF or term frequency is the frequency a term appears in a given document. The term frequency of term t in document d can then be defined as equation 14:

$$TF_{t,d} = \frac{n_{t,d}}{|d|} \quad (14)$$

- Where, $n_{t,d}$ -is frequency of the term t in the document d
- $|d|$ the size of the document d

However, using only TF alone do not provide relevant result as all terms are considered to be equally important. However, inverse document frequency shown in Equation 2balance the inflated weight assigned to frequently appearing terms in the documents and increases the weight of terms that only occurs rarely. 15

$$\mathrm{IDF_i} = \log \frac{|D|}{1+|\{j:t_i \in d_i\}|} \quad (15)$$

Where,

- $|D|$ is number of documents in the document set; $|\{j: t_i \in d_i\}|$ -is the number of documents where the term t_i appears.

Finally the TF-IDF (equation 16) will be:

$$\text{TF-IDF} = \mathrm{TF_{t,d}} * \mathrm{IDF_i} = \frac{n_{t,d}}{|D|} * \log \frac{|D|}{1+|\{j:t_i \in d_i\}|} \quad (16)$$

Cosine Similarity

One of the standard way of quantifying the similarity between two documents A and B is to compute the angle that separate the vectors representing the texts which are done using cosine similarity method. The cosine similarity between documents A and B is dot product of the vectors A and B divided by the Multiplication of the magnitude of vector A with magnitude of vector B and it is formalized as follows:

$$\cos(A, B) = \frac{A \,.\, B}{|A|.|B|} \in [0,1] \quad (17)$$

Where,

- A and Bare m-dimensional vectors over the term set$T = \{t1, \ldots, tm\}$. Each dimension represents a term with its weight in the document, which is non-negative.

As a result, the cosine similarity is non-negative and bounded between $[0,1]$.

Jaccard Similarity Coefficient

Jaccard similarity coefficient is a similarity measure that compares the similarity between two feature sets. When it is applied to measuring similarity between pair of texts, , it is defined as the ratio number of words/terms shared by both text over the number of terms in both texts and it is formalized as follows:

$$\text{Jaccard}(\text{A}, \text{B}) = \frac{|\text{A} \cap \text{B}|}{|\text{A} \cup \text{B}|} \quad (18)$$

Euclidean Distance

Euclidean distance is a standard metric for geometrical problems and widely used in clustering problems, including clustering text. It is a distance between two points and can be easily measured through two-dimensional or three-dimensional space. For a given two points, the Euclidean distance approach returns the distance between those points excluding the direction information existing in the vector based methods. Euclidean distance examines the root of square differences between the coordinated of the pairs in the vectors x and y. Given two documents A and B represented by their term vectors $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ respectively, the Euclidean distance of the two documents is defined using equation 6:

$$D_E(A, B) = D_E\left(\overrightarrow{t_a}, \overrightarrow{t_b}\right) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{1/2}, \quad (19)$$

Where,

- the term set is $T = \{t_1, \ldots, t_m\}$
- $w_{t,a}$ term weight value calculated using TF-IDF

3.4.1.2 Edit Distance

String Edit Distance measures the similarity between two strings based on the number of edit operation applied to transform first string into the second. Hamming and Levenshtein are among the most know edit distance algorithms. Edit distance definition and number and type of actions depend on the algorithm being chosen to calculate the edit distance.

The Hamming Distance takes two strings of equal length and calculates the number of positions at the places where the characters are different [**35**]. It calculates the least number of substitutions needed to transform one string into another. Hamming is mostly used in error-correcting codes in

the fields like telecommunication, cryptography and coding theory. It finds out where the difference is within the two strings.

On the other hand, is Levenshtein Edit Distance allows not only substitution but also insertion and elimination of characters. Levenshtein is perfect to run for finding the similarity on small strings [**35**].

3.4.2 Semantic Text Similarity Measure

Unlike the classical text similarity approaches, Semantic Similarity methods compute similarity using concepts extracted from the text to computing the similarity between concepts. It uses semantic knowledge - ontology during computation.

3.4.2.1 Semantic Knowledge

A semantic knowledge is a semantic network which is composed of a collection of nodes representing concepts and edge representing a semantic relationship between the concepts.

WordNet is a publicly available lexical database developed by Princeton University [**8**]. In WordNet 3.0, there are 206941 words across 117659 SynSets. SynSet represents collection of synonyms words/terms that describes the same fact. Each SynSets has a gloss that defines the concept of the word. Each of these SynSets relates to other SynSets in various semantic relations. WordNet divides the lexicon into five categories also knows as Part Of Speech (POS): Nouns, Verbs, Adjectives, Adverbs and Function verbs.

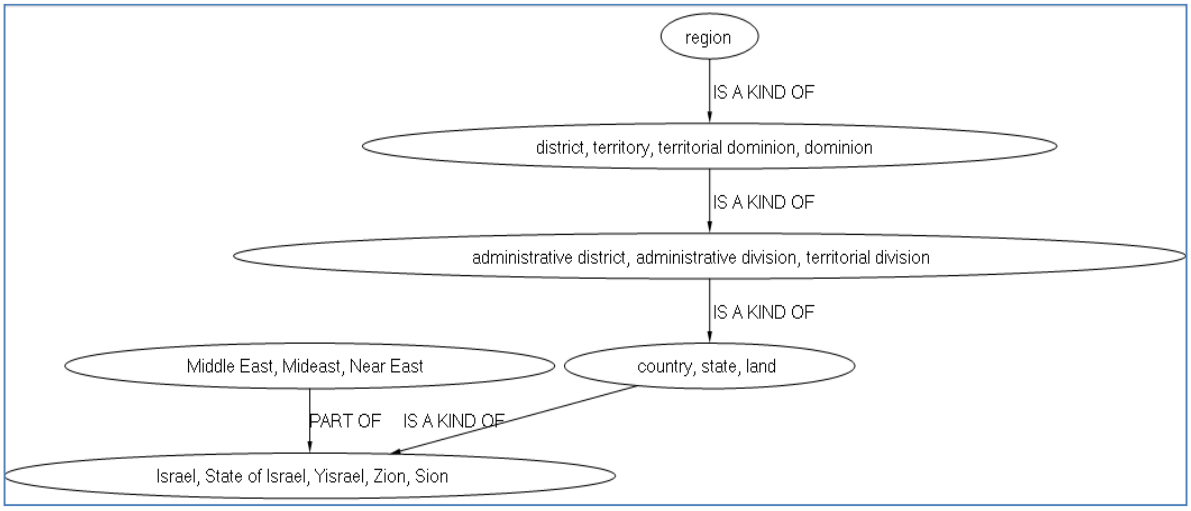

Figure 3-1Fragment of WordNet taxonomy generated by our prototype tool

Semantic Relations

Concepts in the semantic knowledge are related with semantic relations. These relations are represented as a pointer or edge connecting the concepts in the given network.

The most popular and known semantic relations that exist in WordNet knowledge-base are:

- Synonym (≡): two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made or semantically identical. Example in Figure: Israel, State of Israel, Yisrael, Zion, and Sion are synonyms.
- Antonym (Ω): The antonym of an expression is its negation. Antonym of a word x is sometimes not-x, but this definition cannot be generalized. Antonymy is a lexical relation between word forms and it is symmetric. Example: black and White or whiteness.
- Hyponym (≺): x is said to be a hyponymy of y if logically is true and accept when the sentence constructed as "An x is a (kind of) y." It can also be identified as the subordination relation. Example in: Israel is a hyponymy of country, state, land.
- Hypernym (≻) x is said to be a Hypernym of y if logicallyis true and accept when the sentence constructed as "An x has a (has kind of) y." It can also be identified as the super-ordination relation. Example in **Figure 3-1**: country, state and land with Israel.
  Note that, if Y is a Hypernym of X if every X is a Y or Y is a Hyponym of X if every Y is a X. This means if we take **Figure 3-1 Error! Reference source not found.**as an example: concept "Israel" which is Hyponym of "country" and "state" concepts itself is the Hypernym of any country "Israel".
- Meronym (≪): x is said to be a Meronym of y if logically is true and accept when the sentence constructed as "An x is part (Member Of, Substance Of, Component Of etc. of) y." It can also be identified as the part-whole relation, and is generally known as PartOf (also etc.) relation. Example in **Figure 3-1** "Israel: is Part of "Middle East, Mideast, and Near East".
- Holonym (≫): x is said to be a Hypernym of y if logically is true and accept when the sentence constructed as "An x Has Part (HasMember, HasSubstance, HasComponent, etc. of) y." For example **Figure 3-1**: "Middle East" Has Part (a country) called "Israel".

Note that, Y is a Holonym of X if X is part of Y or Y is a Meronym of X. This refers to things being part of others. For example; “Israel” is a Meronym of “Middle East” and “Soma Mideast” is Holonym of “Israel”.

3.4.2.2 Semantic Similarity Measures

Semantic similarity approach can be categorized into two - Distance-based approaches and Content-based approaches. These approaches give a very high-level indication on how early researchers approached the problem. However, the authors in [**36**] proposes a new and better solution called Maximum Enclosure similarity, based on the ratio of the number of shared concepts in the global semantic neighborhood of each concept and the cardinality of the global semantic neighborhood of the second concept.

In the next three sub-sections, we assess the semantic similarity approaches that are categorized into three: distance-based, information content-based approaches and Enclosure Similarity.

3.4.2.2.1 Distance-based Similarity Approaches

The distance based approaches measure similarity between concepts based on the length of the paths between the two concepts in the semantic network. The shorter the distance, the more related the two concepts are.

Distance is usually obtained from the graph-like structure Knowledge base, like WordNet semantic dictionary. Rada & Bicknell [**37**], Leacock and Chodorow [**38**] and Wu and Palmer [**39**]are the most common distance-based algorithms.

The easiest similarity measure is proposed by Rada & Bicknell [**37**]. The similarity between ConceptsC_1 and $C_2$is obtained counting the minimum number of edges between the concepts.

$$Sim_{Rada}(C_1, C_2) = Length(C_1, C_2) \qquad (20)$$

Where, - $Length(C_1,\ C_2)$ is the length between$C_1,\ C_2$.

For example, if we measure similarity concepts1 and concept 2 in Figure$Sim_{Rada}(\text{Ethiopia, Somalia}) = length(Ethiopia, Somalia) = 3$.

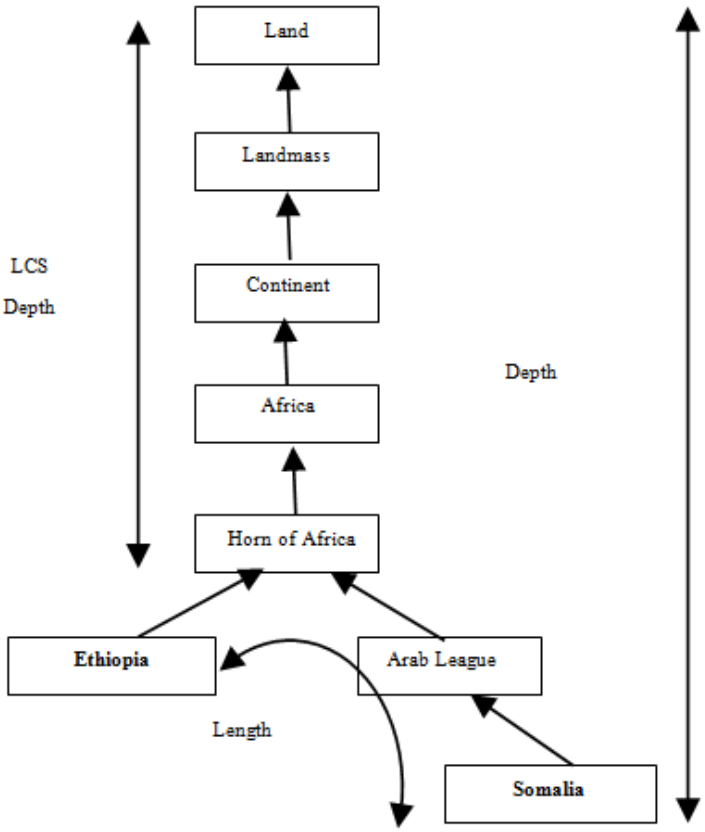

Figure 3-2 Path-based approach shows LCS, depth of the taxonomy and length

Leacock and Chodorow [**38**] is another approach and it is obtained by measuring the shortest path between the two concepts (using node-counting) and the maximum depth of the taxonomy. The similarity of two concepts C_1 and C_2 is obtained:

$$Sim_{LC}(C_1, C_2) = -log \frac{Length}{2*Depth} \quad (21)$$

Where:

- Length is the length of paths between the concepts and
- Depth: the longest depth to the concepts

For example in Figure 3-2

$$Sim_{LC}(Ethiopia, Somalia) = -log \frac{3}{2 \times 6} = \frac{1}{4} = 0.6$$

Wu and Palmer also use the idea of depth in their equation. The Wu & Palmer [**39**] calculates similarity by considering the depths of the two concepts, along with the depth of the LCS (Lowest Common Subsume) where the concepts are related with the hypernyms semantic relationship type. The Wu and Palmer similarity between concepts $C_1 and C_2$ is obtained:

$$Sim_{WP}(C_1,\ C_2)\ () = \frac{2\times LCSDepth}{Length+2\times LCSDepth} \quad (22)$$

Where, Length is the length of the path between the concepts

- Depth is the longest depth to the concepts
- LCS Depth is the depth to the lowest common subsume.

For example, based on Figure 3-2:

$$Sim_{WP}(Ethiopia, Somalia) = \frac{2\times 4}{3+2\times 4} = 0.72.$$

3.4.2.2.2 Information-based Approaches

Information-based method takes into account how much information the two words or concepts share. The more information the two words or concepts share, the more similar the two words are. For example, WordNet is a tree structure and similarity can be measured using the number of nodes that both words share.

In Resnik [**40**] the similarity score of two concepts in an IS-A taxonomy equals the information content value of their lowest common subsume (LCS). And it is donated as:

$$Sim_{Res}(C_1,\ C_2,) = IC(w_{LCS}) \quad (23)$$

Where, $IC(w_{LCS})$ is the information content of their LCS of $C_1 and C_2$ and Information content (IC) of a concept C is computed as negative log likelihood using the probability theory.

Jiang and Conrath extended Resnik's idea to include the distance between the concepts and the lowest common subsumer [**41**]. The similarity value returned by the Jiang and Conrath measure is:

$$JC(C_1, C_2) = IC(C_1) + IC(C_2) - 2 * IC(C_{LCS}) \quad (24)$$

Where:

- IC(x) is the information content of x.

Lin begins with Jiang's hypothesis and uses the universal measure from information theory instead [**42**]. The similarity value returned by the Lin measure is:

$$Lin(w_1, w_2) = \frac{IC(w_{LCS})}{IC(w_1)+IC(w_2)} \quad (25)$$

Where:

- IC(x) is the information content of x.

All three of these measures use the same idea, which is adding the probability of a word appearing in a corpus. The difference is in the equations. An algorithm depending on the amount of shared information indicates an information-based approach.

3.4.2.2.3 Enclosure Similarity

The above Distance and Information based semantic based similarity approaches are only consider the Is-A semantic relation. Getahun et.al [**5**] introduce a new similarity approach called Enclosure similarity. Enclosure similarity considers synonymy, hyponymy and meronymy (IS-A, Part-Of, Member-Of, Instance-Of and Substance-Of) semantic relations.

The concept similarity measure proposed in [**5**] is based on the function of the number of shared and different concepts and it considers their global semantic neighborhoods.

Given a Knowledge Base KB and a threshold, the authors defined the Global semantic neighborhoods of a concept C_i is defined $\overline{N_{KB,\epsilon}}(C_i)$as "set of concepts generated from semantic neighborhood defined with the synonymy, hyponymy and meronymy semantic relations altogether within the same threshold".

It is formalized as follows:

$$\overline{N_{KB,d}}(C_i) = \bigcup_{r\in\{\equiv,\prec,\ll\}} N_{KB}^{r}\,(C_i) \quad (26)$$

Where:

- d is threshold value,
- r is semantic relation $r \in \{Synonyms(\equiv),\ Hyponumy \prec andMeronymy \ll\}$ semantic relations.

Note that, a threshold value is the number of path length separating two concepts in Knowledge base.

Based on [**5**], semantic based similarity/relatedness between two texts can be measured as follows:

Given a text T (i.e., phrase, sentence, etc.), its Concept Set denoted as

$$CS(T) = \{C_1, C_1, \dots, C_m\}$$

Where:

- C_i represents a concept related to at least a word in T

Concept C_i is obtained by after applying text pre-processing techniques such as stop words removal, stemming, tokenization (c.f section text pre-processing 3.2).

Therefore the Text T, which contains m concepts can be described as the follows:

$$T = \{C_1, C_1, \dots, C_m\} \quad (27)$$

Given two texts T_1 and T_2 and their corresponding Concept Set CS (T_1) and CS (T_2), vector space model is used to represent distinct concepts as axis, together with their associated weighted score w_i of each text T_i in an n-dimensional space.

$$\vec{V} = (\langle C_1, w_1\rangle, \langle C_2, w_2\rangle, \dots, \langle C_m, w_m\rangle, \dots, \langle C_n, w_n\rangle) \quad (28)$$

Where w_m is the weight score associated to concept $C_m \in (CS(T_1) \cup CS(T_2))$, $1 \leq m$, n and $n=|CS(T_1) \cup CS(T_2)|$.

The weight of a concept w_m in vector is 1 if concept c_m is member of the Concept Set CS of the other anchor text T_j where j is 1 or 2. The weights are calculated as follows:

$$w_m = \begin{cases} 1 & if\, c_m \in CS(T_2) \\ max(ES(c_i, c_j)) & otherwise \end{cases} \quad (29)$$

Where, ES refers to the Enclosure Similarity, which is computed as the ratio of shared concepts of the global semantic neighborhood of the concepts over the number of concepts in the neighborhood of the second concept. It is formalized as follows: text:

$$ES(c_i, c_j) = \frac{|\overline{N_{KB,\epsilon}}(C_i) \cap \overline{N_{KB,\epsilon}}(C_j)|}{|\overline{N_{KB,\epsilon}}(C_j)|} \quad (30)$$

After the weights are computed the similarity between the two texts $T_i$and T_j is calculated by the following equation:

$$SemRel(NT_1, NT_2) = Cos(\overrightarrow{V_1}, \overrightarrow{V_2}) = \frac{\overrightarrow{V_1}.\overrightarrow{V_2}}{|\overrightarrow{V_1}|\times|\overrightarrow{V_2}|} \in [0,1] \quad (31)$$

Its result is the dot product of the vectors V_i and V_j over the product of the magnitude of each vector.

3.5 Xml Document Similarity Measure

XML is used mostly to express and exchange data among enterprise applications usually machine to machine communication. It has become the standard language for data transmission and exchanging on the Internet. As a matter of fact, most web-based applications deal with web data by translating them into XML format. Recently most of commercial database systems (Oracle, IBM DB2) provide tools to deliver information in XML format and to store XML data [**43**]. Moreover, due to the increase number of XML documents on the Web, there is an increasing need to automatically process those structurally rich documents for information retrieval, similarity clustering, and search applications.

Since, News Feed items are XML files, an interesting approach in relation to this thesis work is, to efficiently store and retrieve XML documents is based on grouping together similar XML documents. Therefore, a good management of XML content has become a main research issue in order to get relevant and unduplicated information [**43**]. Among those researches, measuring similarity between XML documents has been broadly studied; for example in text document searching, document clustering, copy or plagiarism detection, text document retrieval, filtering, categorization [**43**], and News Item similarity [**5**].

It is common to measure the similarity between XML documents by considering their structures (structure-based approach), contents (content based approach) or utilize both structure and content information which is called Hybrid-based approach.

3.5.1 Structured-based

According to [**44**,**45**], each XML document has both a logical and a physical structure. Logically, the document is composed of declarations, elements, comments, character references, and

instructions. These components are indicated in the document have their own markup. Physically, the document is composed of units called entities. An entity may refer to other entities to cause their contribution in the document. A document begins in a root or document entity.

Measuring structural similarity between XML documents has become a key component in various applications, including XML mining, schema matching, and web service discovery, News Feed clustering and merging, etc. Structure of XML documents has many models such as: tree based model, map based model, path based model and so on. The main purpose of XML structure analysis is to measure the structure similarity of XML documents [**46**]. According to [**47**] and [**48**], the following are among the most common structure similarity algorithms:

- **Tree Edit Distance (TED) Similarity**: Measures the minimum number of edit operation need to transform the structure of one XML document into the other.
- **Tag similarity**: Measures how closely the set of tags match between two XML documents. Documents that use a similar set of tags will likely have a similar schema. The tag sets of the two documents can be compared to measure their overlap.
- **Edge matching**: It combines the simple node (tag) matching technique by estimating similarity between two XML documents based on their matching nodes.
- **Path similarity**: XML documents are compared with respect to their corresponding sets of paths: the more paths two XML documents share in common, the more similar they are.
- **Fourier Transform**: It computes the similarity by extracting the sequence of start tags and end tags from the documents, and convert the tag sequence to a sequence of numbers to represent the structure of the documents. The number sequence is then viewed as a time series and Fourier transform is applied to convert the data into a set of frequencies. The similarity between two documents is computed in the frequency domain by taking the difference in magnitudes of the two signals.
- **XML/DTD similarity**: it is a method for measuring the structural similarity between an XML document and a DTD (Document Type Definition) grammar by taking into account the level (i.e. depth), in which the elements occur in the hierarchical structure of the XML and DTD tree representations. Elements at higher levels are considered more relevant, in the comparison process, than those at lower levels.

3.5.2 Content-based

The content of the XML documents refers to the textual data [**49**]. In content based approach, the text view of the XML documents is main source for measurement, without assigning any special significance to the tags or the structural information. Previously content based can be measured using the classical text based similarity measure and recently semantic similarity based content was introduced.

3.5.3 Hybrid

Even though the contents of the XML document play major role in classifying XML documents, it has limitations in distinguish the differences involved in the structure of the XML documents. Recently, researchers showed the importance of combining structure and content [**7**], [**50**], [**7**] and [**51**] in measuring similarity value. Therefore, a hybrid approach for XML documents combines the structural and content similarity values.

In recent years, there have been some works on integrating semantic and structural similarity in the XML comparison process. For instance, authors of [**52**] introduce a combined structural/semantic XML similarity approach integrating IR semantic similarity assessment in a traditional Edit Distance algorithm. They consider the various semantic relations encompassed in a given reference taxonomy/ontology (e.g. WordNet) while comparing XML documents.

In [**53**] tag similarity (synonyms and stems) is used instead of tag syntactic equality for measuring XML documents. In [**52**] the authors proposed an integrated semantic and structure based XML similarity approach, taking into account the semantic meaning of XML element/attribute labels in XML document comparison.

3.6 Summary

Text-based similarity measure is about measuring two or more strings with each other to find out how similar they are. It is categorized as classical text similarity measure and semantic based similarity measure. Classical text similarity may include Vector Space Model and Edit Distance. On the other hand semantic based similarity measure uses Knowledge Base during computation. One of the most common semantic networks is WordNet, which is composed of collection of

concepts also called SynSet related with semantic relations (such as: Synonym, Hyponym, Hypernym, Meronym, etc…).

Semantic similarity measures can be categorized into Distance-based, content based and Enclosure similarity approaches. Distance-based approaches measure the length of the path between two words; Information-based method takes into account how much information the two words or concepts share within a restricted IS-A semantic relationship; Enclosure similarity takes into account several semantic relations Is-A, Synonym, Hyponym and Meronym together.

As RSS/Atom feed formats are XML file format, the different kinds of XML similarity measurement methods discussed in Section 3.2 have been used to measure similarity between two News items.

CHAPTER FOUR
LINK-BASED RSS NEWS ITEMS SIMILARITY MEASURE

4.1 Introduction

In this Chapter, we introduce the Link-based similarity measurement framework shown in Figure 4-1 Our Framework is composed of three main and interacting components: Related link set extractor, Link Similarity evaluator, and Text Similarity evaluator.

The approach accepts two news items as input and extracts the link sub-element as entry point to the actual news.

The related link set extractor component accepts a web page as input and returns the set of anchor elements/ links (outgoing links) citied in that page. This component returns only those anchor elements which are relevant and related to the original page.

Link Similarity component computes the similarity between two links. The similarity is computed after identifying the concepts existing in the URL and anchor text. Once the concepts are identified semantic based text similarity between the links is computed using the text similarity method proposed in [5].

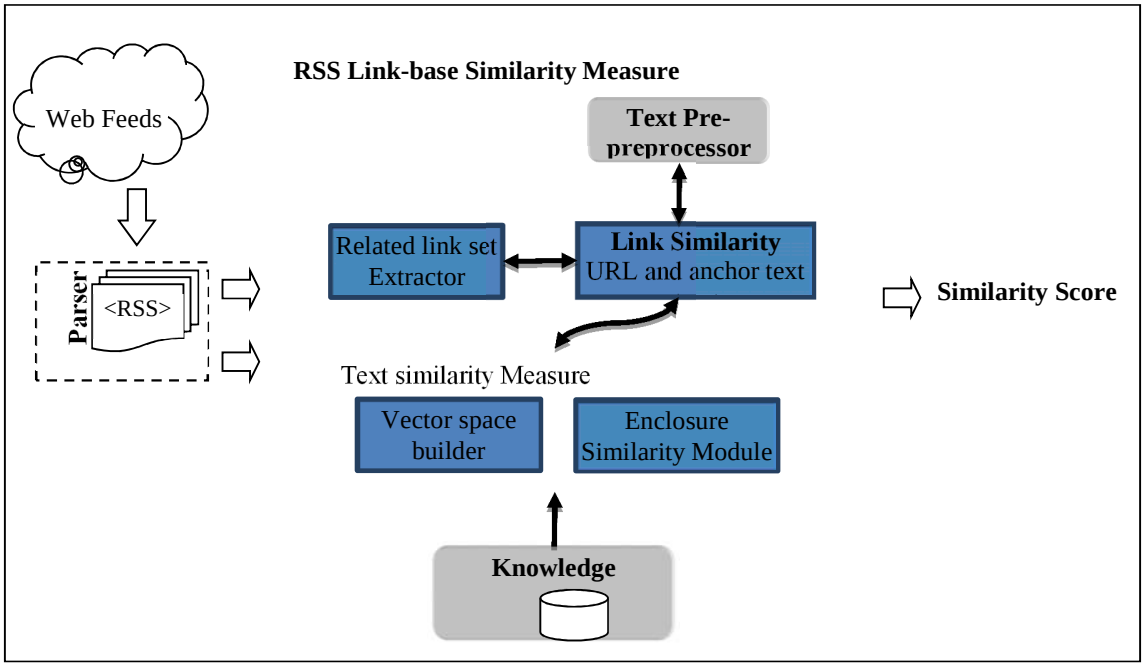

Figure 4-1: Overview of link based news feed similarity measure framework

4.2 **Preliminaries**

In addition to news feed link sub-element, our similarity measure also considers related news links of each news page, at different depths of the web graph. In hyperlink analysis, researches represent the web as graph in which nodes represent pages and edges corresponds to hyperlinks connecting nodes of the graph [**13**], [**12**], [**54**], and [**55**].

Adopting this principle, we use the link sub-element of each input news articles as the root node of the graph and determine its degree of similarity with another link sub-element of another News article by extracting related links at different depth of the web graph.

4.2.1 **Rooted Ordered Tree**

A rooted ordered tree T is a set of nodes denoted as $\{\boldsymbol{r}, \boldsymbol{n_i}\}$, where r is the root and $\boldsymbol{n_i}(\boldsymbol{i} = \mathbf{1} \ldots \boldsymbol{m})$ are ordered elements and immediate children of r, and each represent an element/sub-tree of r rooted at$\boldsymbol{n_i}$.

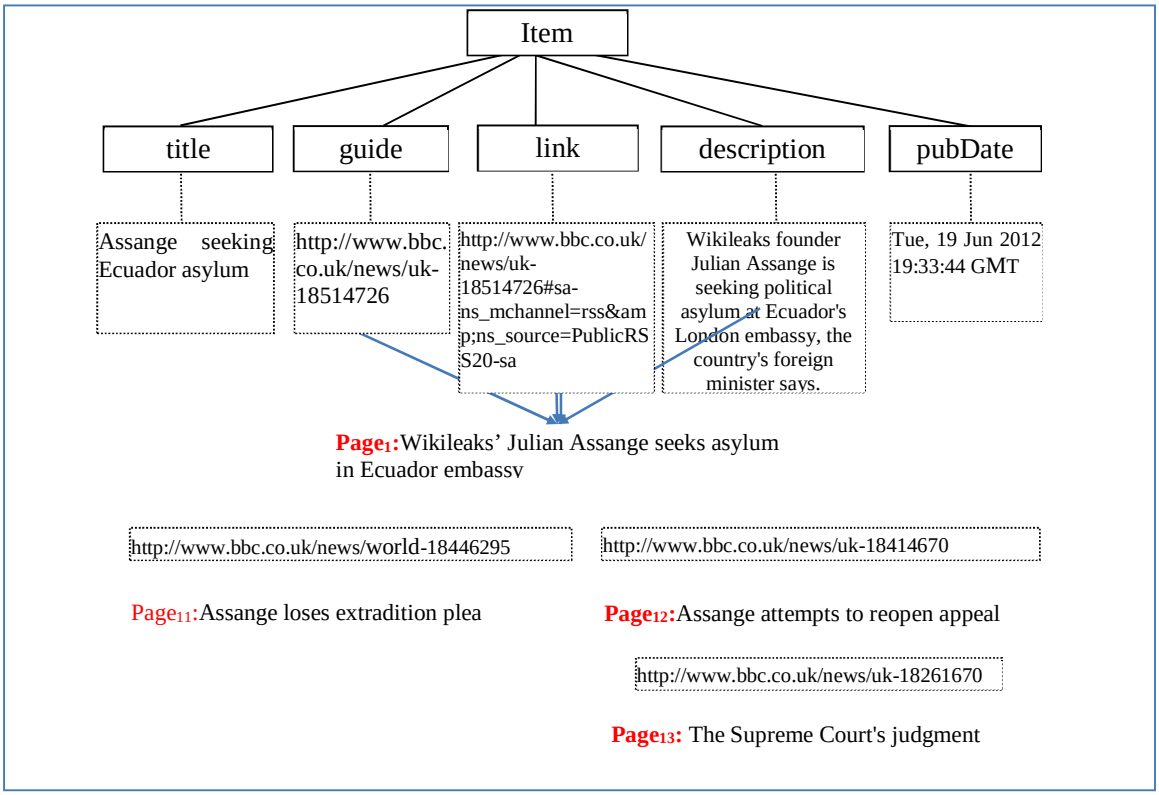

Figure4-2 Ordered rooted tree

News providers publish news items enclosed in a channel C. A news item I_i in C is a short summary of a news page P_i whose address is stored in sub-element of I_i called link. It is likely that the page P_i stores address of set of related news pages.

Figure4-2 shows ordered label tree representing sample RSS news items extracted from BBC. This news item contains the address of the actual source news page ($Page_1$) in the link element. The $Page_1$ contains the hyperlink to $Page_{11}$ and $Page_{12}$ as related and $Page_{12}$ contains hyperlink to $Page_{13}$.

4.2.2 Link Graph

A graph G is a data structure having three components: a vertex set V(G), an edge set E(G), and a relation that associated to each edge connecting two vertices called its endpoints[**56**].

Each edge is an ordered pair of nodes (u, v) representing a directed connection from u to v, if and only if page u contains a hyperlink to page v. We call this directed graph of the link graph G. In directed graph the out-degree of a node u is the number of distinct paths $(u, v_1) \dots (u, v_K)$ (i.e., the number of links from u), and the in-degree is the number of distinct paths $(v_1, u) \dots (v_k, u)$ (i.e., the number of links to u).

A path is a sequence of distinctive vertices connected by edges. A path from node u to node v is a sequence of paths $(u, u_1), (u_1, u_2), \dots (u_k, v)$.

On the other hand an undirected graph is the same with directed graph, except that, there is an edge between u and v if there is a link between u and$v$, without regard to whether the link points from u to v or the other way around. The degree of a node u is the number of edges incident to$u$.

4.3 News Feed Link Sub-Elements

A news feed exists in either RSS or Atom format and each has different versions. These two feed format have different structures caused by the use of different tag names in representing related information. However, the sub-element <link> is common in both versions

The content of this element is used as entry point to access the actual news containing set of anchor elements. The anchor element in the news contains the address/hyperlink of pages which are cited as they are related to the container news.

4.4 URL and Anchor Concept Set Extractor

Anchor element in an html contains hyperlink/URL that contains address of a resource it is pointing to textual information that semantically describes the resource.

The URL in a news page is used to refer to another news page, news section, news article, person, or perhaps a location profile. Related links inside News page are sometimes called a recommendation links suggested by the news writer and they attract the user to get more information about that specific clicked News. Therefore, in addition to the link element, these outgoing links together with their corresponding anchor text information is considered in our similarity measure. The URL contains set of concepts that describe the information embedded in the path and similarly the anchor text contains information that describes the resource.

Online News providers such as BBC, ABC News and CNN usually store related News in the same subdirectory.

For example, http://www.bbc.co.uk/news/world-us-canada-18128995.shtml contains about directories where the information is stored (News) and additional information that describe the content of the resource pointed by the link (world-us-canada). Thus, there is a need to extract concepts embedded in URL and anchor text as well. Algorithm 4-1 is accept a text representing either URL or anchor text as input and returns list of concepts extracted from it. The algorithm pre-process the text (ignoring stop words – it include domain name, URL file name extensions, numerical values, separating characters or symbols, , less relevant words, etc); map each term in the pre-process text into concept representing it referring the knowledge base KB.

	Algorithm 4-1: Concept Extractor
1	Input: Str : String // describe either URL or anchor text
2	Variable: C: Concept terms: set // list of words
3	Output: Concepts: Set //set of concepts extracted from the text
4	Begin
5	terms= Pre-process (str)
6	For each w in terms
7	C = MapWord2Concept (w, KB) // map the word into the concept that contains the it
8	Concept.add(C)// add the concept to concept set
9	Next
10	Return concept
11	End

4.5 Related Link Set Extractor

In addition to news feed link sub-element, our similarity measure also considers related news links of each news page, at different depths of the web graph. In web hyperlink analysis area, most researches have been using links by considering the web to be a graph data structure, where pages form nodes and hyperlinks form edges between the nodes of the graph [**13**], [**12**], [**54**], and [**55**]. Therefore, we treat the link sub-element of each input news articles as a graph root node and determine its degree of similarity with another link sub-element of another News article by extracting related links at different depth of the web graph.

Therefore, after obtaining each Link item and construct the Link graph it is easy to study the links and compute the similarity. In our case each node contains URL and anchor text information together.

Given two News Feeds their link sub-elements will be the root link node of their respective graph. Traverse through the link graph top to bottom to obtain set of links which are connected to the root links either directly or indirectly. These links are extracted following the next steps:

1. Start from the root of the graph, visit the page pointed by the edge and extract the content of the web page using Pasternack and Roth algorithm [**57**].
2. Exclude large amount of less informative and typically unrelated material such as navigation menus, forms, user comments, and advertisements.
3. Extract only links and their anchor text inside the HTML document.
4. Extract recommended and related links together with their anchor text information. Notice that there is neither standard followed by web site designers nor fixed position that that contains related links.
5. Apply training data in order to trace container elements such as and extract the textual concepts from each link set and their anchor text.

Therefore, we use the term level (depth) to represent different sets of nodes:

- Nodes in level-zero only contain the starting points corresponding to root links which is the URL of the link sub-element of the RSS/Atom news feed.
- Nodes in level-one are nodes cited by level-zero nodes only.
- Nodes in level-two are nodes cited by nodes in level-one, but not by nodes in level-zero and so on.

Note that, in terms of edges, we say that edge e is in level-n if one endpoint of e is in level-n and the other endpoint is in level n-1.

4.6 **Link and Semantic Based Similarity Measure**

Our similarity score values for any two given News feeds NF_1 and NF_2 reflect how closely the two News feeds are similar based on their link element information. In this work, in order to measure the similarity between URLs and anchor texts we adopt the text similarity measure of [**5**].

Definition1: [Hyperlink Element]

A link or hyperlink is a simple element[6] containing the address/URL of an outgoing page and also an optional anchor text.

Example: Human rights lawyer Moon sets South Korean presidential bid

The content of the anchor element is the content of the href attribute (i.e., URL text "places/south-korea" describing places in South Korea) and the anchor text that summarizes the page pointed by the URL.

Definition 2: [Item Similarity]

Given two news items I_i and I_j, the similarity in between is computed as the average similarity of their corresponding sub-elements.

Definition 3: [Text Similarity]

Given two strings T_1 and T_2, their similarity denoted as TextSimilarity (T_1, T_2) is computed using the cosine of the vectors representing each text.

To each text, the vector contains set of distinct terms/words with associated weight. The weight of a term in a text is computed using enclosure similarity measure proposed in [**5**].

Definition 4: [Hyperlink Similarity]

Given two outgoing hyperlinks p_i and p_k, their similarity denoted as Sim(p_i, p_k) is computed as the weighted sum of text similarity between their corresponding link and anchor elements. It is formalized as:

$$Sim(p_i, p_k) = \alpha TextSimilarity(p_i.link, p_k.link) + (1-\alpha) \times TextSimilarity(p_i.anchor, p_k.anchor) \quad (32)$$

Where,

[6] An XML element without a child is simple element otherwise it is complex or composite element

- $\alpha \in [0,1]$is a weight value reflecting the importance of link in comparison to anchor.
- p_i. link returns the content of the hyperlink element
- p_i. anchor returns the anchor text of the hyperlink

4.7 Link-based Item Similarity

One of the known approaches in computing similarity between complex elements is aggregating the similarity between sub-elements enclosed in it which are computed aggregating the similarity between their corresponding components – tag names and textual contents.

In this work, we compute the similarity by aggregating the similarity between link elements at different levels.

Definition 5: **[Link Similarity]**

Given two news items I_i and I_j, each represented as a graph, and depth d, the link similarity denoted as LinkSim is computed following the steps presented in Algorithm 4-2 and formalized in equation 33:

$$Linksim(i_i, i_j, d) = \begin{cases} TextSimilarity(i_i.link, i_j.link) & d = 0 \\ \frac{\sum_{l=1}^{n}\sum_{k=1}^{m} \mathrm{sim}(p_l,p_k)}{\mathrm{count}} & \mathrm{d} = 1 \\ \frac{\sum_{l=1}^{d} \mathrm{linksim}(i_i,i_j,l)}{\mathrm{d}} & \mathrm{d} > 1 \end{cases} \quad (33)$$

The algorithm accepts two news items and a depth d as inputs. At a depth of Zero, only link elements of the items are used to compute the similarity and it comes down to measuring the text similarity between link elements as shown in line 9to 11.

Otherwise, before computing the link similarity all outgoing hyperlink or link of each news items at depth d is extracted. The outgoing links are set of related news items cited in the source page at a given depth d which are extracted using the maximum segmentation algorithm of Pasternack and Roth[**57**] (as shown in Line 12-13 and detailed in Section 4.4).

At the Depth (level) of one, the link similarity is computed as the average of link similarity between outgoing of news item l_i with each outgoing hyperlink of the news item$l_j$as shown in lines14-21. Otherwise, the similarity is computed as the aggregate of the link similarity at depth run between 1 and d as shown in line 23-27.

	Algorithm 4-2: Link Similarity
	Input:
1.	i_i, i_j: item
2.	d : integer
3.	KB : Semantic Knowledge
	Variables:
4.	n, m: integer
5.	LinkSimSum: Decimal
6.	Count: Integer
7.	$P_i[]$, $P_j[]$: String //contains outgoing links
	Output:
8.	LinkSim: Decimal
	Begin
9.	If(d==0)then
10.	LinkSim = $\text{TextSimilarity}(i_i.\text{link}, i_j.\text{link})$
11.	Else
12.	$P_i[]$=ExtractRelatedOutgoingLink(i_i, d)
	// n is the no of such links at level d
13.	$P_j[]$= ExtractRelatedOutgoingLink(i_j, d)
	// m is the no of such links at level d
14.	If (d==1)then
15.	For l=1 to n
16.	For k =1 to m
17.	LinkSimSum += sim(Pi[l], Pj[k])
18.	If (LinkSimSum>0) count ++
19.	Next
20.	Next
21.	LinkSim = LinkSimSum / count
22.	Else
23.	For j=1 to d

24.	LinkSimSum += LinkSim(I_i, I_j,j)
25.	Next
26.	Linksim = LinkSimSum / d
27.	End If
28.	End If
29.	Return Linksim
	End

Definition 6: [Combined Item Similarity]

The combined similarity between items is computed as the average similarity between news items computed using title and description sub-elements (as presented in [**5**]) and the link similarity computed using **Algorithm 4-2**. It is formalized as follows:

$$\mathrm{ComSemRel}(i_1, i_2) = \frac{\mathrm{ItemSimilarity}(i_1,i_2)+\mathrm{Linksim}(i_i,i_j,d)}{2} \quad (34)$$

Where Item Similarity is computed following the approach presented [**5**] as the average similarity between their corresponding sub-elements as formalized as follows:

$$\mathrm{ItemSem}(i_1, i_2) = \frac{\mathrm{ElementSimilarity}(e_i,e_j)}{|i_1|\times|i_2|} \forall\, e_i \in i_1\ \forall e_i \in i_2 \quad (35)$$

4.8 Computational Complexity

The computational complexity of our LinkSim algorithms are identified using the worst case analysis and using the RAM machine. The computation is dependent on the number of links in each source (X and Y) and the time complexity of computing text similarity. It comes down to the following basic parts:

$$O(\mathrm{LinkSim}) = O\big(x \times y \times O(\mathrm{TextSimilarity})\big)$$
$$= O(x \times y \times n \times m \times n_c^2 \times d^4)$$

Where

- X is number of out-going links in the first source- page
- Y is the number of out-going links in the second source - page
- $O(\mathrm{TextSimilarity})$- the worst case time complexity needed to compute text similarity.

According to [**36**], the time complexity of text similarity measure depends on the number of concepts extracted from the two texts, n and m, the maximum number of words per a concept, n_c, and the depth, d , in the knowledge Base. It is denoted as: $O(TR) = O(n \times m \times n_c^2 \times d^4)$

Therefore, the complexity of computing link similarity depends highly on the text similarity and link sets of each links

4.9 **Summary**

In this chapter we have presented a link-based similarity measure for a dedicated to news feeds. Our measure considers only link sub element of news feed XML file and its related outgoing links. It is computed first by extracting related hyperlink from a news page, and then it constructs the link graph for each News feed; finally it applies the Semantic Relatedness similarity measure. The proposed measure can be easily computed by specifying the Level (depth) of the graph, the threshold value between concepts.

CHAPTER FIVE
PROTOTYPE & EXPERIMENTATION

5.1 Introduction

In this Chapter, we present the set of experiments conducted to validate our proposed approach. The rest of the chapter is organized as follows. In Section 5.1, we present development tool and techniques. In Section 5.2, we present the user interface of our prototype. In Section 5.3, we present the dataset and text pre-processing techniques used in the experiment. Section 5.4 presents Evaluation methods followed by experimental results in Section 5.5. Finally, we summarize the Chapter in Section 5.6.

5.2 Architecture of the Prototype

The tool developed to validate our proposal is a desktop application developed with C#. Our tool uses the desktop RSS aggregator, RSSOwl, to download news feed from the net and use it as source for our RSS news items similarity approach.

The main components of the link similarity tool are:

- Knowledge Base– a WordNet2.1[7] lexical taxonomy, exploited in evaluating text similarity. We have extended the WordNet.Net [**58**], an open-source .NET framework library for WordNet, developed by .Malcolm Crowe and Troy Simpson.
- RSSOwl[8]is a Microsoft Windows desktop RSS aggregator for managing (add, edit, remove) feeds from any source. It is used to save selected information in various formats for offline viewing and sharing. We use RSSOwl to collect news feeds for our experiment.

[7] http://wordnetcode.princeton.edu/2.1/

[8] http://www.rssowl.org/

- The Similarity component is responsible to compute the similarity a pair of texts, hyperlinks or links. It measures similarity after

i) Pre-processing text values

ii) generating a vector for each text,

iii) computing the similarity between words/concepts, texts, using the Enclosure Semantic Measure component

iv) computing the similarity between items

2. Graphical User interface component allows a user to visualize the WordNet knowledge base in graphical format, enter two sample hyperlinks to be compared; enter inputs for computing semantic similarity between words, texts, hyperlinks, and items. A sample query interface is discussed in Section 5.3.

5.3 User Interface

The user interface developed for our system has four tabs – semantic neighborhood, word similarity, text similarity and link based items similarity respectively. These tabs help the user to interact with the different similarity computation services. Each tab provides the user with options of selecting threshold value and viewing the final score of the request.

Error! Reference source not found. shows the semantic neighborhood of 'Arab' within with a threshold value of '4'. The nodes in the figure represent concepts and the edges represent the semantic relationship existing between the content nodes. The concept of Arab is related with other concepts with Is Kind Of and Member Of semantic relationships.

Error! Reference source not found.The second tab computes enclosure similarity between pair of words/terms or concepts and displays their similarity score. Figure 5-1 shows the similarity score of car and ambulance at threshold value =1.

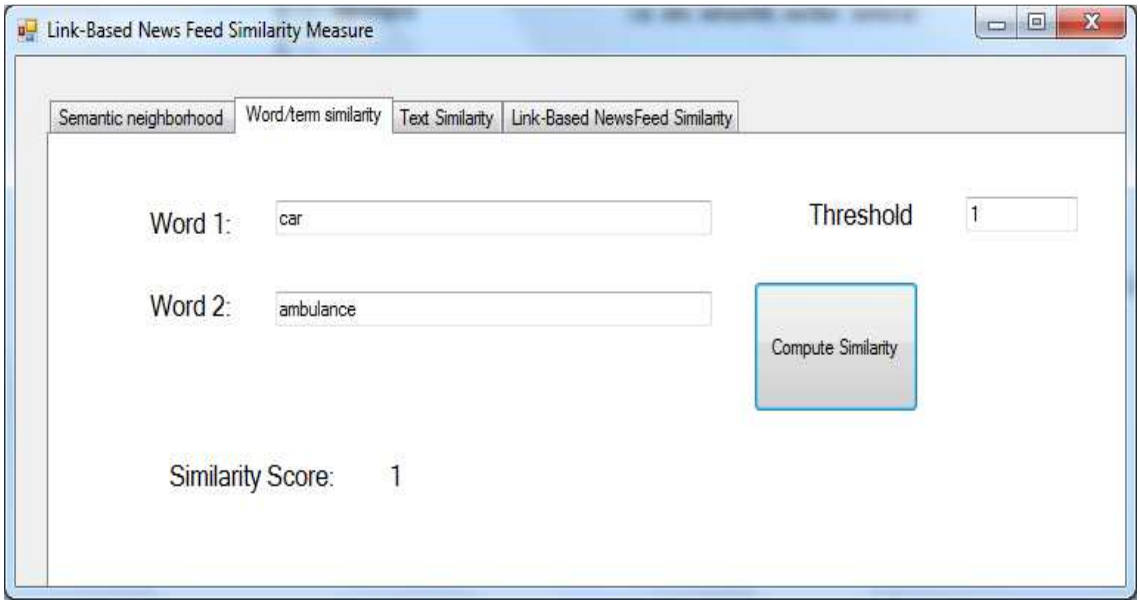

Figure 5-1 Enclosure similarity of Word/Term or concept

The third tab computes semantic similarity of any two given texts using Enclosure similarity and displays their similarity score. Figure 5-2shows the similarity score between news title “Egypt Elections: Second Round of Parliamentary Vote” and “Egyptians Vote in New Round of Parliamentary Elections” at a threshold value of “1”. The table in the figure shows the vector space generated for each text with associated weight value.

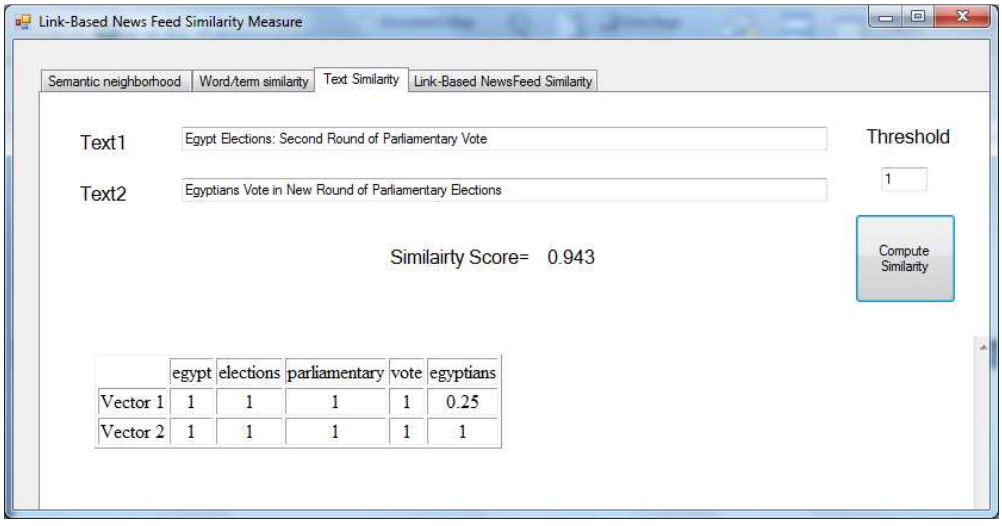

Figure 5-2 Text similarity

The fourth tab has an interface that enables the user to compute the similarity between pair of news feeds using their title, description and link elements as content descriptor. Finally it shows a similarity scores value.

For example: Figure 5-4shows two two news feeds extracted from BBC and Press Release(PR)Web[9] Sports. The similarity in between is computed using their “title”, “description”, “link” elements as descriptor within a threshold of 2 and the result is shown in.

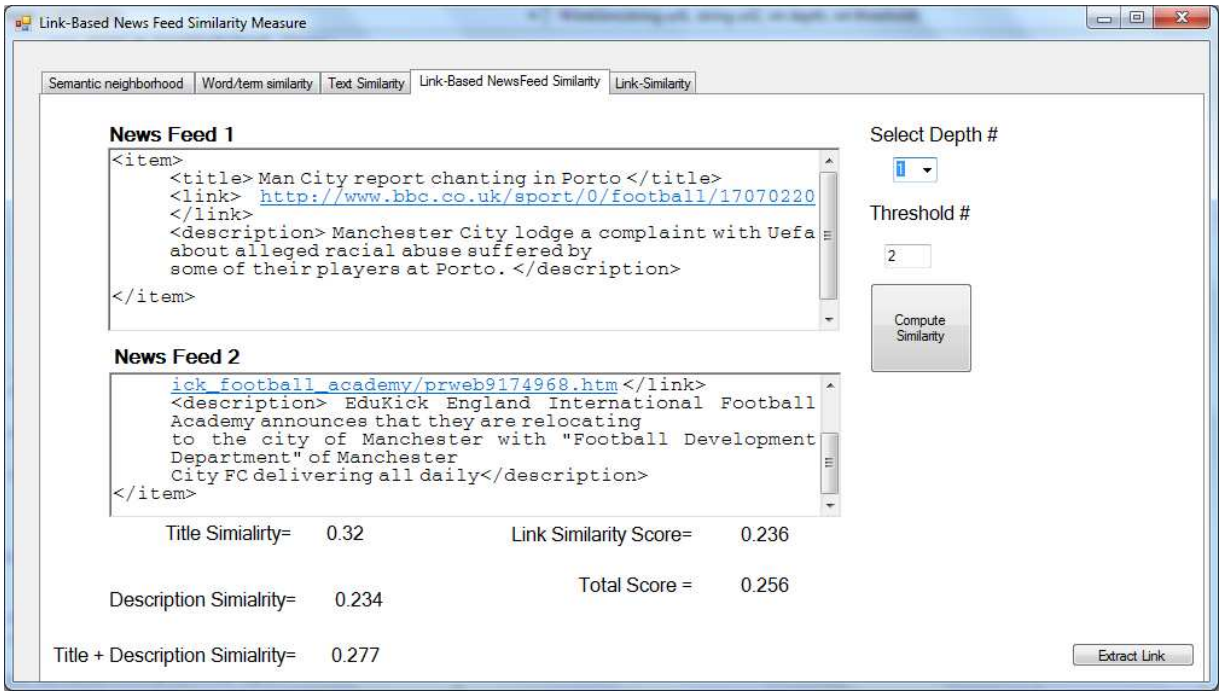

Figure 5-3 News feed similarity measure for a given two news feeds from PRweb and BBC

5.4 Data Collection and Experimentation

Data is collected from different News sources using RSSOwl. For our similarity measures, we created collection of News feeds. In this collection we select 50 News Feed pairs. These paired news feeds are selected randomly and they may relate or unrelated.

[9]www.**prweb**.com/

5.4.1 Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. It transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Before modeling the content of the input text, a cleaning step is performed to remove stop-words, and stem available words using C# code includes different codes for text preprocessing such us:

- Trim- remove white space before and after the concept;
- URL validation- check for valid URL and return texts after the main word
- Number and symbol (wild cards characters) remover- since we are only concerned with words and terms, we remove numbers and wild cards from the given text.
- Stemmer(Porter's stemmer algorithm) [**31**]
- Stop-word removal taken from [**59**] and a complete file can be found in [**60**].

During examining and performing pre-experimentation tasks we found some interesting problems that may decrease our experimentation result.

While we perform text pre-processing techniques, many interesting concepts were found. For example, the lack of a complete stemming algorithm to compute a term "dies" and "flies". It returns "di" and "fli" instead of "die" and "fly" respectively.

Secondly, Because of the related news links extracted from pages such as CNN website are generated automatically, it is difficult to retrieve the HTML source code online for related stories. As a result, our tool can't able to read them. This would decrease the percentage accuracy of the link similarity score for depth different from zero. Finally, we found that WordNet is not complete enough in describing all terms and concepts. For instance, the Meronymies of Ethiopia are "Addis Ababa, New Flower, capital of Ethiopia, Lake Tana and Lake Tsana". However, there are a lot of place in Ethiopia that are not included in the dictionary. As a result of this the similarity value might goes down.

5.5 Evaluation Method

In this Section, we describe our evaluation of the similarity performance. The objective of the set of experiments conducted in this Section is to show only the effectiveness of the approach. To

measure effectiveness, we show first how link item of a News feed is effective in measuring similarity. To evaluate our approach, we have developed a test method and built a test environment.

The scores we are going to analyze are as follows:

- News Similarity computation using only the content of title sub-element. News feed titles usually contain the most important information about the actual content.
- News Similarity computation using only description/summary sub-element News feed description or summary usually contain the first sentences of the actual content.
- News Similarity computation using only link sub-element News feed link usually contain a URL to the actual content. We also consider news feed similarity based on their outgoing related links at a depth of one and two.
- News Similarity computation using combination of title, description and link sub-elements at a given depth zero and one.

5.6 **Experiment Results and Discussions**

Based on previously proposed method, we run set of experiments on 50 pair of news feeds and computed the similarity score using title, description, $link_{Sim}$, combination of title and description ($Item_{Sim}$), "title, description, link" ($titleDescLink_{Sim}$)and level one link similarity (Level1Link) Then, we compared the average similarity for these six sub-element similarity scores. The average of the similarity value is shown in Table 5-1.

Table 5-1Average semantic similarity score for different sub-elements of news feed

	$title_{sim}$	$description_{sim}$	$Link_{Sim}$	$Item_{Sim}$	$titleDescLink_{Sim}$	$Level1_{Link}$
Average Similarity Score	0.099	0.101	0.191	0.086	0.126	0.168

The table shows that link similarity ($Link_{sim}$) achieves a 0.023 improvement over Level1Link similarity, 0.065 over $Item_{sim}$, 0.090over$titleDescLink_{sim}$, 0.092over$description_{sim}$and 0.105over$title_{sim}$.

In addition, we have compared the result of our link-based similarity method at depth zero, one, two and three with SimRank and co-citation at similar depth. In Figure 5-4they-axis indicates the average similarity score and x-axis represents similarity level (whole number).Our experiment

result shows that our LinkSim at level zero has better similarity score than SimRank, co-citation and LinkSim at level one and two.

Again, when we look at graph in Figure 5-4 we found that, the similarity score for our proposed LinkSim algorithm is decreases when the number of link depth increases. This is so as the depth increases the number of concepts extracted from the anchor increases. In spite of this, LinkSim provide a better result when the level decreases i.e., level zero and one. Specially at level zero, we found similarity score using LinkSim, while co-citation and SimRank score zero.

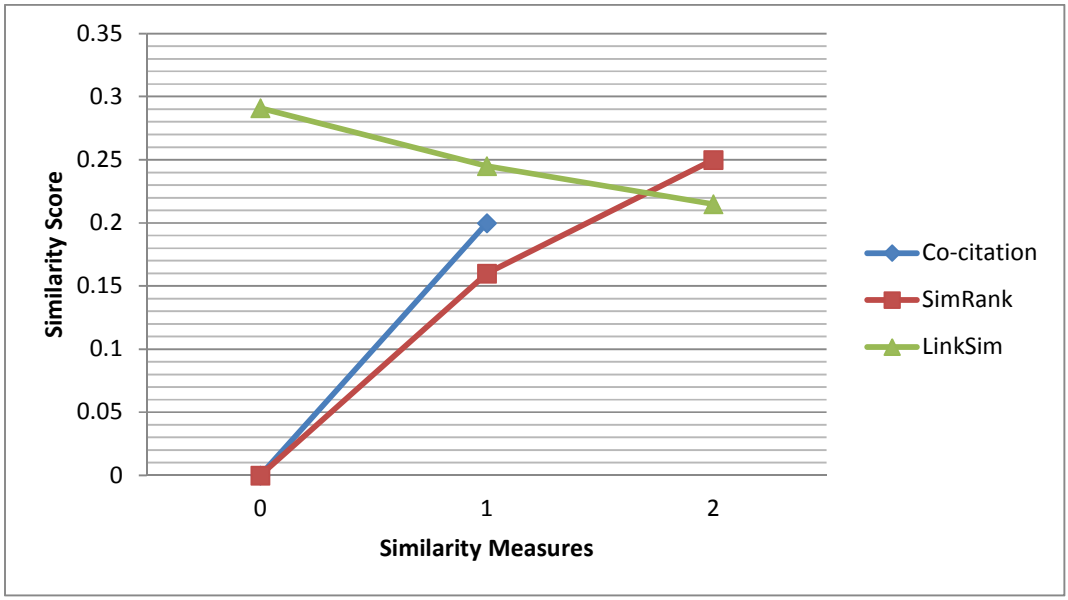

Figure 5-4 News feed link based similarity scores

5.7 **Summary**

In this Chapter, we have presented our link bases news feed similarity measure prototype which is designed to present the semantic neighborhood of a concept within a given threshold. It computes enclosure similarity between any two concepts and compute the link based similarity between news items. This chapter also details our experimentation techniques and results for selected 50 news feed pairs. The experimental results shows that proposed solution is scores better result in comparison with SimRank and co-citation.

CHAPTER SIX
CONCLUSION

6.1 Conclusion

In this thesis we address the on how to measure similarity between two News items. We have given a theoretical overview of some of the other text based similarity methods.

We introduced a link based semantic similarity measure and defined link information text based semantic similarity measures, which used for measuring contextual similarity. We also integrate the link bases similarity with title and description sub-elements for butter similarity score.

In addition to this, internal and external information of a News Feed can be combined effectively using our similarity measure method. That means our method not only suits for News item file similarity measure only, but also is applicable for utilizing external information such as Recommended related things by the News provider in the News page and any related links inside the article.

We conducted several experiments on sample News feed formats with different sub elements of the News feeds Item. Based on our analysis, we saw that the proposed similarity measures, which use the link text information, have potential for measuring which News stories would be less similar.

6.2 Future Works

Considering the novelty of the arguments treated in this thesis, the work done constitutes only the starting point of a more wide research line. Indeed many improvements and open points need to be solved.

The further work of this study might involve improving Link scores, testing more in-depth the Link extraction and perhaps constricting the similarity matrix. Specially, extracting related links and their anchor text at a given depth need further study. Currently, we are using first extract the whole News article from the HTML document, extract links from the extracted content and extract the rest links by finding the recommended related link position. However, we could have an

algorithm for extracting the related link of a News page at a given depth. This might increase the computation speed considerably.

On the other hand some similarity measures require word senses rather than tokens as input data. These word senses (word meanings or concepts) are primitive semantic constituents of a document. A word can have multiple senses: The word “bank” has the following senses

- A financial institution (“He transferred his money to another bank”) or
- A slope next to a river (“He sat on the bank of the river”)

Depending on the context the similarity measures required to use the most suitable word sense, which is the case in automatic sense disambiguation.

BIBLIOGRAPHY

[1] Danny A. and Andrew W, *Beginning RSS and Atom programming*.: Wiley Publishing, 2005.

[2] A. Danny and W. Andrew, *Beginning RSS and Atom programming*. Indiana, Canada: Wiley Publishing, 2005.

[3] Dave J., "RSS and Atom in Action," *Manning Publications*, 2006.

[4] Dave J., *RSS and Atom in Action.*: Manning Publications, 2006.

[5] Joe, T., Chbeir, R., Marco, V., and Kokou, Y., Fekade G., "Relating RSS News/Items," *ICWE*, pp. 442-452, 2009.

[6] N., Zhang, D., Yu, Y., and Duan, J., Zhang, "An improved method for classifying document based on Structure and Content," *Academy Publisher AP-PROC-CS-10CN007*, 2010.

[7] Richard C. and Kokou Y., Joe T., "A Hybrid Approach for XML Similarity," *LE2I Laboratory UMR-CNRS*.

[8] George A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41.

[9] Günter Ewald, "Wadsworth Publishing.," in *Geometry: An Introduction*., 1971, pp. 106, 181.

[10] D., & Markman, A.B. Gentner, "Structural alignment in analogy and similarity," *American Psychologist*, vol. 52 (1), pp. 45–56, 1997.

[11] Monika Henzinger, "Link Analysis in Web Information Retrieval," *ICDE Bulletin*, vol. 23, Sept 2000.

[12] J. Han, and Y. Sun, P. Zhao, "P-Rank: A Comprehensive Structural Similarity Measure over Information Networks," *In proc. of Int'l. Conf. on Information and Knowledge Management*, pp. 553-562, 2009.

[13] Z.,Michael,R.,King, L. Lin, "PageSim: A Novel Linkbased Measure of Web Page Similarity," 2006.

[14] Fernando F. Souza, Valéria C. Times, and Fabrício Benevenuto Paulo R. Santos, "Towards integrating Online Social Networks and Business Intelligenc," *In Proceedings of the IADIS International Conference on Web Based Communities and Social Media (WBC'12). Lisbon, Portugal*, 2012.

[15] W. H., Lancaster, J., Paradesi, M. S. R., & Weninger, T. Hsu, "Structural link analysis from user profiles and friends networks: A feature construction approach," *Proceedings of ICWSM*, pp. 75-80, 2007.

[16] H.G. Small, "Co-citation in the scientific literature: A new measure of relationship between two documents," *Journal of the American Society for Information Science*, vol. 24(4), pp. 265–269, 1973.

[17] M.M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14(1), pp. 10–25 , 1963.

[18] Dubes RC Jain AK, "Algorithms for clustering data," *Prentice-Hall, Inc.*, 1988.

[19] G., Widom, J. Jeh, "SimRank: A measure of structural-context similarity," *SIGKDD*, pp. 538-543, 2002.

[20] R. Amsler, "Application of citation-based automatic classification.," *Unpublished manuscript:The University of Texas at Austin, Linguistics Research Center, Austin, TX.*, 1972.

[21] M. R. Lyu, and I. King. Z. Lin, "MatchSim: A novel neighbor-based similarity measure with maximum neighborhood matching," *In CIKM'09: Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pp. 1613–1616, 2009.

[22] D. Fogaras and B. R´acz, "Scaling link-based similarity search," *In WWW'05: Proceedings of the 14th international conference on World Wide Web ACM*, pp. 641–650, 2005.

[23] S., Kim,S.,Park, S. Yoon, "C-Rank: A Link-based Similarity Measure for Scientific Literature Databases," 2011.

[24] Brin S., Motwani R., and Winograd T. Page L., "The PageRank citation ranking: Bringing order to the web," *Technical report, Stanford University*, 1998.

[25] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46(5), pp. 604–632, 1999.

[26] J. Dean and M.R. Henzinger, "Finding Related Web Pages in the World Wide Web," *Proc. Eighth Int'l World Wide Web ConfElsevier Science,New York,* , pp. 389-401, 1999.

[27] J., & Henzinger, M.R. Dean, "Finding related pages in the world wide web.," *Computer Networks*, vol. 31(11–16), pp. 1467–1479, 1999.

[28] Z.,Song,R.,Nie,J., and Wen, J. Dou, "Using Anchor Texts with Their Hyperlink Structure for Web Search," July 2009.

[29] N, Kevin S. Eiron, "Analysis of Anchor Text for Web Search," *SIGIR*, 2003.

[30] L.W., and Chen, S.M. Lee, "New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents," *IEA/AEI*, vol. 4031, pp. pp.1280-1289., 2006.

[31] Porter M., "An algorithm for suffix stripping Program," vol. 14(3), pp. 130-137, 1980.

[32] Y., Witten, I.H., and Wang, D., Wen, "Token Identification Using HMM and PPM Models.," *Intelligence, Proceedings of the 16th Australian Conference on Artificial Intelligence*, vol. 2903, pp. 173-185, 2003.

[33] A. Wong, and C.-S. Yang Salton, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18(11), pp. 613–620, November 1975.

[34] G., & Buckley, C. Salton, "Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management," vol. 24(5), pp. 513-523, 1988.

[35] ed Paul E. Black, ""Levenshtein distance", in Dictionary of Algorithms and Data Structures," *CRC Press LLC*, 1999.

[36] Fekade Getahun TADDESSE, "Semantic-aware News Feeds Management Framework," *Laboratoire Électronique, Informatique et Image – LE2I*, 2010.

[37] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," *Systems, Man and Cybernetics, IEEE Transactions*, vol. 19, no. 1, pp. 17-30, Jan/Feb 1989.

[38] Claudia Leacock and Martin Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed.: MIT Press, 1998, pp. 305-332.

[39] Zhibiao Wu and Martha Stone Palmer, "Verb Semantics and Lexical Selection," in *Proceedings of the 32th Annual Meeting on Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 133-138.

[40] Philip Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448-453.

[41] Jay J. Jiang and David W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy ," in *In International Conference Research on Computational Linguistics* , 1997, pp. 19-33.

[42] Dekang Lin, "An Information-Theoretic Definition of Similarity," in *In Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296-304.

[43] B. Jeong, D. Lee, H. Cho, and B. Kulvatunyou, "A Novel Approach to Measuring Structural Similarity between XML Documents," 2006.

[44] Lee D., Cho H., Kulvatunyou B. Jeong B., "Novel Approach to Measuring Structural Similarity between XML Documents," 2006.

[45] KIM W., "XML document similarity measure in terms of the structure and contents," *2nd WSEAS Int*, pp. 205-212, 2008.

[46] Dongzhan Z., Ye Y. JiangjiaoAn D. Na Z., "an improved method for classifying XML documents based on structure and content," pp. 426-430, 2010.

[47] David Buttler, "A Short Survey of Document Structure Similarity Algorithms".

[48] Richard Chbeir, and Kokou Yetongnon Joe Tekli, "An overview on XML similarity: background, current trends and future directions".

[49] Y. Ng I. Garcia, "Eliminating Redundant and Less-Informative RSS News Articles Based on Word Similarity and a Fuzzy Equivalence Relation," *ICTAI* , pp. 465-473, 2006.

[50] Kevin R., "Indexing XML Documents: A Hybrid Approach," 2003.

[51] GT., Nayak,R., Bruza,P. Tran, "Combining Structure and Content Similarities for XML Document Clustering," 2008.

[52] R. Chbeir and K. Yetongnon. J. Tekli, "Semantic and Structure based XML Similarity: An Integrated Approach," *In Proceedings of the 13th Interventional Conference on Management of Data (COMAD'06)*, pp. 32- 43, 2006.

[53] G. Guerrini and M. Mesiti. E. Bertino, "A Matching Algorithm for Measuring the Structural Similarity between an XML Documents and a DTD and its Applications," *Elsevier Computer Science*, vol. 23-46, 200.

[54] Liu B. Marzolla M., "Linka Analysis and Web Search".

[55] M. Thelwall, "Exploring the link structure of the Web with network diagrams," *Journal of Information Science* , vol. 27(6) , pp. 393-402, 2001.

[56] F. Harary, "Graph Theory," *Addison Wesley*, 1975.

[57] Roth., D. Pasternack. J., "Extracting Article Text from the Web with Maximum Subsequence Segmentation," *International World Wide Web Conference Committee (IW3C2)*, pp. 971-980, 2009.

[58] Crowe M. Simpson T., "WordNet.Net," *http://opensource.ebswift.com/WordNet.Net*, 2005.

[59] G. Salton, "The SMART Retrieval System—Experiments in Automatic Document Processing.," *Prentice-Hall, Inc., Upper Saddle River, NJ,* , 1971.

[60] The stop-words list. [Online]. http://members.unine.ch/jacques.savoy/clef/englishST.txt

[61] J.A., Higgins, D., Soglasnova, S. Goldsmith, "Automatic Language-Specific Stemming in Information Retrieval," *in: CLEF '00 Revised Papers. Presented at the Workshop of Cross-Language Evaluation, Forum on Cross-Language Information Retrieval and Evaluation.*, 2000.

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

This thesis has been submitted for examination with my approval as university advisor.

_____

Advisor