STRUCTURING FREE-FORM TAGGING IN ONLINE NEWS

Cher Han (Andy), LAU

Submitted in fulfilment of the requirements for the degree of
Master of Information Technology (Research)
School of Information Technology
Faculty of Science and Technology
Queensland University of Technology
2009
Keywords

Folksonomy, social tagging, online news, digital news
Abstract

Tagging has become one of the key activities in next generation websites which allow users selecting short labels to annotate, manage, and share multimedia information such as photos, videos and bookmarks. Tagging does not require users any prior training before participating in the annotation activities as they can freely choose any terms which best represent the semantic of contents without worrying about any formal structure or ontology. However, the practice of free-form tagging can lead to several problems, such as synonymy, polysemy and ambiguity, which potentially increase the complexity of managing the tags and retrieving information. To solve these problems, this research aims to construct a lightweight indexing scheme to structure tags by identifying and disambiguating the meaning of terms and construct a knowledge base or dictionary.

News has been chosen as the primary domain of application to demonstrate the benefits of using structured tags for managing the rapidly changing and dynamic nature of news information. One of the main outcomes of this work is an automatically constructed vocabulary that defines the meaning of each named entity tag, which can be extracted from a news article (including person, location and organisation), based on experts suggestions from major search engines and the knowledge from public database such as Wikipedia. To demonstrate the potential applications of the vocabulary, we have used it to provide more functionalities in an online news website, including topic-based news reading, intuitive tagging, clipping and sharing of interesting news, as well as news filtering or searching based on named entity tags.

The evaluation results on the impact of disambiguating tags have shown that the vocabulary can help to significantly improve news searching performance. The preliminary results from our user study have demonstrated that users can benefit from the additional functionalities on the news websites as they are able to retrieve more relevant news, clip and share news with friends and families effectively.
# Table of Contents

Keywords ................................................................................................................................. i  
Abstract ................................................................................................................................... ii  
Table of Contents ................................................................................................................... iii  
List of Figures .......................................................................................................................... vi  
List of Tables ............................................................................................................................ viii  
List of Abbreviations ............................................................................................................... ix  
Statement of Original Authorship .......................................................................................... x  
Acknowledgments ..................................................................................................................... xi  

## CHAPTER 1: INTRODUCTION ......................................................................................... 1  
1.1 Executive Summary ......................................................................................................... 1  
1.2 Background ....................................................................................................................... 2  
1.3 Objectives ......................................................................................................................... 3  
1.4 Significance ....................................................................................................................... 4  
1.5 Thesis Outline ................................................................................................................... 5  

## CHAPTER 2: ONLINE NEWS ........................................................................................... 7  
2.1 Introduction ...................................................................................................................... 7  
2.2 News Tags ......................................................................................................................... 8  
2.3 News Characteristics ....................................................................................................... 9  
2.4 User Tagging in Social News ......................................................................................... 11  
2.5 News Aggregator ........................................................................................................... 12  
2.5.1 Google News ............................................................................................................... 12  
2.5.2 Digg.com ..................................................................................................................... 12  
2.5.3 Newsvine ..................................................................................................................... 12  
2.5.4 Personalised news delivery ....................................................................................... 13  
2.6 Named Entity Indexing ................................................................................................. 13  
2.7 Word Sense Disambiguation ......................................................................................... 14  
2.8 Knowledge Base ............................................................................................................. 16  
2.9 Limitations of Current Systems .................................................................................... 18  
2.9.1 Lack of reading time support ................................................................................... 18  
2.9.2 Support for Named Entity based Search ................................................................ 19  
2.9.3 Organisation tools ..................................................................................................... 19  
2.9.4 Content management ............................................................................................... 20  
2.9.5 Scattered social activity ............................................................................................ 20  
2.10 Conclusion ...................................................................................................................... 21  

## CHAPTER 3: FOLKSONOMY ......................................................................................... 22  
3.1 Introduction ...................................................................................................................... 22  
3.2 Tagging Practices .......................................................................................................... 23  
3.2.1 Public tagging ............................................................................................................. 23  
3.2.2 Creator tagging .......................................................................................................... 23  
3.2.3 Machine tagging ........................................................................................................ 24  
3.3 Tagging Motivation ........................................................................................................ 24
# Structuring Free-Form Tagging in Online News

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Types of Folksonomy</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>Types of Tagging System</td>
<td>29</td>
</tr>
<tr>
<td>3.6</td>
<td>Benefits of Folksonomy</td>
<td>30</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Low overhead cost</td>
<td>30</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Finite result</td>
<td>31</td>
</tr>
<tr>
<td>3.6.3</td>
<td>Social context sensitive</td>
<td>32</td>
</tr>
<tr>
<td>3.6.4</td>
<td>Easy implementation</td>
<td>32</td>
</tr>
<tr>
<td>3.7</td>
<td>Social Tagging disadvantages</td>
<td>33</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Low precision</td>
<td>33</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Low recall</td>
<td>33</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Lack of control</td>
<td>34</td>
</tr>
<tr>
<td>3.8</td>
<td>Tag structuring</td>
<td>34</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Lightweight approach</td>
<td>35</td>
</tr>
<tr>
<td>3.8.2</td>
<td>Heavy weight approach</td>
<td>37</td>
</tr>
<tr>
<td>3.9</td>
<td>User Study On Tagging</td>
<td>39</td>
</tr>
<tr>
<td>3.10</td>
<td>Tags Visualisation</td>
<td>40</td>
</tr>
<tr>
<td>3.11</td>
<td>Social Tagging Feature</td>
<td>43</td>
</tr>
<tr>
<td>3.12</td>
<td>Conclusion</td>
<td>45</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>47</td>
</tr>
<tr>
<td>4.2</td>
<td>News Indexing</td>
<td>47</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Indexer</td>
<td>47</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Entity extractor</td>
<td>47</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Named Entity Mapper</td>
<td>47</td>
</tr>
<tr>
<td>4.3</td>
<td>News System</td>
<td>48</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Graphical User Interface</td>
<td>48</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Visualisation</td>
<td>48</td>
</tr>
<tr>
<td>4.3.3</td>
<td>User management</td>
<td>48</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Organisation</td>
<td>49</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Retrieval</td>
<td>49</td>
</tr>
<tr>
<td>4.4</td>
<td>Extended Functionality</td>
<td>50</td>
</tr>
<tr>
<td>5.1</td>
<td>Design</td>
<td>52</td>
</tr>
<tr>
<td>5.2</td>
<td>Named Entity Extraction</td>
<td>53</td>
</tr>
<tr>
<td>5.3</td>
<td>Knowledge base</td>
<td>54</td>
</tr>
<tr>
<td>5.4</td>
<td>Expert Suggestion</td>
<td>55</td>
</tr>
<tr>
<td>5.5</td>
<td>Sense Mapping</td>
<td>56</td>
</tr>
<tr>
<td>5.6</td>
<td>Dictionary</td>
<td>60</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Polysemy</td>
<td>60</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Synonymy</td>
<td>61</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Presentation</td>
<td>62</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusion</td>
<td>64</td>
</tr>
<tr>
<td>6.1</td>
<td>User Scenarios</td>
<td>66</td>
</tr>
<tr>
<td>6.2</td>
<td>System Details</td>
<td>67</td>
</tr>
<tr>
<td>6.3</td>
<td>Functionality</td>
<td>68</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Reading tools</td>
<td>69</td>
</tr>
</tbody>
</table>
6.3.2 Support for named entity based search ................................................................. 71
6.3.3 Tagging ................................................................................................................ 72
6.3.4 Event Follow up .................................................................................................. 73
6.3.5 Integrated social support ..................................................................................... 74
6.3.6 Visualisation ........................................................................................................ 75
6.4 Conclusion .............................................................................................................. 76

CHAPTER 7: RESULTS AND DISCUSSION ................................................................. 80
7.1 Scope ..................................................................................................................... 79
7.2 Dataset ............................................................................................................... 80
7.3 Dictionary ............................................................................................................ 80
7.4 Mapping Accuracy ............................................................................................... 82
7.5 Tag Convergence ................................................................................................ 85
7.6 Retrieval Performance ........................................................................................ 88
7.7 Users Study .......................................................................................................... 89
7.8 Discussion ............................................................................................................ 91

CHAPTER 8: CONCLUSION AND FUTURE WORK .................................................. 93
8.1 Summary of achievements .................................................................................... 93
8.2 Evaluation Performance ....................................................................................... 93
8.3 Trends and future work ....................................................................................... 95

BIBLIOGRAPHY ......................................................................................................... 97
List of Figures

Figure 2.1. Visualised news characteristics ................................................................. 10
Figure 2.2. User tags distribution .............................................................................. 11
Figure 3.1. General tagging motivation ..................................................................... 26
Figure 3.2. Tagging motivation in photo sharing ....................................................... 27
Figure 3.3. Broad folksonomy ([74]) ......................................................................... 27
Figure 3.4. Narrow folksonomy ([74]) ....................................................................... 28
Figure 3.5. Long tail of folksonomy .......................................................................... 29
Figure 3.6. Folksonomy hierarchy ............................................................................. 29
Figure 3.7. Cognitive process of tagging [77] ............................................................. 31
Figure 3.8. Cognitive process of categorisation [77] .................................................... 31
Figure 3.9. Tag index (from ABC) .............................................................................. 41
Figure 3.10. Flag tag list (from Mixx) ....................................................................... 41
Figure 3.11. Tag cloud (from del.icio.us) ................................................................. 41
Figure 4.1. System architecture ............................................................................... 49
Figure 4.2. User function in reading news ................................................................. 50
Figure 5.1. Design of the indexing process ............................................................... 53
Figure 5.2. Calais named entity extraction and classification model ....................... 54
Figure 5.3. Paradigm of tag sense mapping ............................................................... 58
Figure 5.4. Sense disambiguation process flowchart ............................................... 59
Figure 5.5. Sample RDF entry snapshot ................................................................... 62
Figure 6.1. Main Graphical User Interface ............................................................... 68
Figure 6.2. Related news and related entity panel .................................................... 69
Figure 6.3 News reading without highlighting ......................................................... 70
Figure 6.4. News reading with highlighting Wikipedia abstract ................................ 70
Figure 6.5. Search criteria formulation ..................................................................... 71
Figure 6.6. Search result for news article ................................................................. 71
Figure 6.7 Search result for entity ........................................................................... 71
Figure 6.8. Tagging toolbox ..................................................................................... 73
Figure 6.9. Viewing clipbooks .................................................................................. 73
Figure 6.10. Current news sharing model ................................................................. 74
Figure 6.11. OneStopNews sharing model ............................................................... 75
Figure 6.12. Tag cloud ............................................................................................. 75
Figure 6.13. User activity log ................................................................................... 76
Figure 7.1. Graphical representation of ‘Australia’ RDF after enrichment ............... 81
Figure 7.2. Related entities of “Kevin Rudd” at different time ................................. 82
Figure 7.3. Mapping performance result ................................................................. 84
Figure 7.4. Tag space growth ................................................................................................................ 86
Figure 7.5. Tag space size comparison .................................................................................................. 86
Figure 7.6. User study result ................................................................................................................. 90
Figure 7.7. Average rating per question ................................................................................................ 91
List of Tables

Table 2.1. Examples of Wikipedia redirection ................................................................. 17
Table 2.2. Wikipedia disambiguation of “NBA” ................................................................. 18
Table 2.3. Online news website used in the study ............................................................ 18
Table 3.1. Golder and Huberman tagging feature .............................................................. 25
Table 3.2 Guidelines for combining Web 2.0 and Semantic Web ........................................ 36
Table 3.3. Ontology properties for folksonomy formalisation ............................................. 38
Table 5.1. Spelling corrections example ............................................................................ 53
Table 5.2. Named entity tags and their expert suggested meaning Wikipedia title ................ 56
Table 5.3. Example of polysemous terms ......................................................................... 61
Table 5.4. Example of synonyms ...................................................................................... 61
Table 5.5. Field used in the dictionary ............................................................................. 63
Table 6.1. GUI Components ............................................................................................. 68
Table 7.1. Statistics of two dictionary instances ................................................................. 81
Table 7.2. Example of correctly mapped tags .................................................................... 83
Table 7.3 Example of incorrectly mapped tags .................................................................. 83
Table 7.4. Named entity tag result ................................................................................... 83
Table 7.5. Creator tag result ............................................................................................ 84
Table 7.6 User tag result ................................................................................................ 84
Table 7.7. Tag convergence performance .......................................................................... 87
Table 7.8. Tag convergence performance (by category) ..................................................... 88
Table 7.9. Information retrieval performance evaluation ................................................... 89
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>.NET</td>
<td>Microsoft .NET Framework</td>
</tr>
<tr>
<td>AJAX</td>
<td>Asynchronous Javascript and XML</td>
</tr>
<tr>
<td>API</td>
<td>Application Programable Interface</td>
</tr>
<tr>
<td>CSS</td>
<td>Cascading Style Sheet</td>
</tr>
<tr>
<td>DC</td>
<td>Dublin Core</td>
</tr>
<tr>
<td>DDC</td>
<td>Dewey Decimal Classification</td>
</tr>
<tr>
<td>KB</td>
<td>Knowledge base</td>
</tr>
<tr>
<td>NE</td>
<td>Named Entity</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RSS</td>
<td>Rich Site Summary</td>
</tr>
<tr>
<td>WSD</td>
<td>Word Sense Disambiguation</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>
Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: _________________________

Date: ___________________________
Acknowledgments

It is a pleasure to thank the many people who made this thesis possible.

It is difficult to overstate my gratitude to my supervisor, Dr. Dian Tjondronegoro. With his enthusiasm, his inspiration, and his great effort during my candidature, he helped me to achieve success in this research. Throughout my thesis writing period, he provided encouragement, advice, teaching and inspirational ideas. This work would not have been possible without him.

I would also like to thank my associate supervisor, Dr. Richi Nayak, who has provided me with precious knowledge in data mining and information retrieval techniques. I would have been lost without her kind assistance on the research design and constructive inputs on thesis writing.

Lastly, and most importantly, I wish to thank my parents for their support and love. To them I dedicate this thesis.
Chapter 1: Introduction

This chapter outlines the executive summary (Section 1.1), background (Section 1.2), the overarching objectives (Section 1.3) and the significance (Section 1.4) of this research study. Finally, it includes an outline of the remaining chapters of the thesis (Section 1.5).

1.1 INTRODUCTION

The usage of Internet technology has grown at an impressive rate over the past decades and various technologies have been developed to provide users with easy accessibility to online information. Online news has emerged as an important Web publishing medium which enables the public to obtain current affairs information from breaking news stories at any time and from anywhere. This can be witnessed by the use of online news websites in the aftermath of the September 11 incident in which the World Trade Centre was attacked [1]. While users can easily gain access to large amounts of news articles using the Internet, the interactivity between news systems and readers is still limited [2]. The main interactive functions supported by current major news websites are still limited to multimedia navigation and basic retrieval, while content organisation features such as tagging and news sharing mechanisms are still lacking [2].

The use of tagging is becoming a widely accepted technology used in emerging web to facilitate information retrieval and collaborative information sharing [3]. Users are allowed to freely select any keywords (as tags) to describe a web resource. The tagging activity and the generated results are defined as folksonomy, which signifies the process of generating taxonomy by the folks’ effort [4]. The success of folksonomy is evident through the implementations in the currently popular social websites for managing media contents such as photos [5], videos [6] and bookmarks [7]. However, the use of folksonomy is still not optimized for major commercial online newspapers [8, 9]. For users, folksonomy system presents a low cognitive cost classification method that does not require them to learn a formal hierarchical taxonomy, which often can limit their choice of terms selection and hinder active involvement and participation in news annotation/sharing. From the system perspective, folksonomy delegates the indexing task to the public and therefore avoids the initial setup cost which is normally required to gather professional cataloguers to design and create the initial classification model, and thus the overall annotation system becomes more dynamic and user context sensitive.
Tagging provides a convenient approach that facilitates resource annotation, and the success of its application in many social websites. Although tagging-based sites implementation has been successful, one barrier preventing users from consistently finding all relevant information is also caused by the use of freely chosen tags. As tagging enables term based search techniques, the performance of the search result varies according to the keyword selected. Different users may utilise different keywords to describe the same content and the same tags may be used to describe different resources of a dissimilar nature. The main challenges to solving this problem is resolving tag ambiguity and explicitly defining the semantic meanings of the tags.

Research studies show that structuring folksonomy is an effective method of overcoming the above mentioned limitations. Folksonomy can be structured using lightweight approaches by interpreting the statistical information from the tag usage ([10], [11]) or applying formal structuring using thesauri, controlled vocabulary and ontology [12]. Controlled vocabulary is widely used in various key knowledge management fields such as library management and cataloguing [13]. It is always considered as a primary solution for tag structuring. A common approach used to construct controlled vocabulary for news articles is to perform named entity indexing [14]. This approach is well suited to the nature of news articles as news articles mainly consist of information about named entity types such as a person, location and organisation [15]. However, to date, the implementation of named entity based tag structuring is still not popular and not yet fully implemented on commercial news websites.

This research develops a framework that utilises the named entity indexing for use on an online news tagging platform. Different types of tags are extracted from the news articles and a lightweight algorithm is developed to index them. The named entity index is then utilised to complement the existing keyword based search and to provide additional support for tagging in the current news system. The tagging functionality aims to provide additional functionality to help the users in managing and sharing news clippings. The outcome of this project will be to effectively enhance the information retrieval performance and provide a better collaborative information sharing experience for the online news consumers.

1.2 BACKGROUND

Online news development has advanced significantly in the past few years. An online news system is deliver multimedia content on the latest current affair at a very short frequency. This causes the online news system to become a major channel that provides users with easy accessibility to obtaining real time information. However, users face the problem of information overload and require an effective mechanism to browse and retrieve relevant news articles. Current news websites utilise traditional keyword based document indexing (such as TF-IDF), which can be improved by complementing them with named entity indexing [14]. However, such
implementations are not yet fully applied in the present systems and have not been delivered to consumers. By enabling named entity based browsing and retrieval, users are allowed to retrieve news articles based on a particular entity (such as person or company) in which they are interested.

Another problem in current online news systems is that users are not given the flexibility to perform additional news reading activities, such as tagging interesting news articles to create news clippings and sharing information with other users who share the same interests. The success of social tagging websites shows that tagging is an effective mechanism for resource organisation and collaborative sharing [16] but we have yet to see such functionality become fully implemented on any commercial news website. Van der Wal first coined the terms “folksonomy” which describes the phenomenon of generating taxonomy from folks’ effort [4]. Studies reveal that folksonomy is useful, but various deficiencies caused by the free form nature exist, such as; polysemy, synonym use and ambiguity which affect the performance of tagging. Structuring folksonomy will significantly improve the system retrieval performance and the usability of tagging.

The main research questions that will be answered in this thesis are:-

1. How to develop a lightweight indexing scheme to index and structure named entity tags in online news?

2. How can the indexing scheme be utilised to structure social tagging and provide management and social functionality to achieve a collaborative online news system?

1.3 OBJECTIVES

The primary outcome of this research project is the development of a lightweight indexing scheme to index and structure named entity tags in news articles in order to produce a dictionary which functions as a controlled vocabulary for use in a news system.

The second outcome in this research is a prototype of an online news system which implements social tagging as an additional tool to provide users with increased functionality in organising and managing news clippings. The lightweight dictionary is constructed as a controlled vocabulary that will strengthen information retrieval and social tagging activities in online news.

The indexing scheme will be tested against a large dataset of news articles and the system will be evaluated based on the retrieval performance using standard information retrieval measurements.
1.4 SIGNIFICANCE

The significance of this research lies equally in the news indexing method and the collaborative news system.

As compared to indexing approaches that require preset rules and regular model training, the proposed indexing scheme adopts a lightweight and dynamic dictionary construction approach. The indexing scheme should be able to accurately map various different types of tags detected in news articles such as named entity tags, creator tags and users tags. A dictionary is automatically constructed as a controlled vocabulary to govern the tags and improve the overall retrieval performance. The dictionary identifies the root form and various surface forms by uniquely mapping each individual tag to its respective node in the knowledge base. The structured dictionary can be linked to other systems to enrich the information of the tags. The dictionary provides users with facilities to: (1) browse and search news articles using named entities, and (2) allow faster tagging using suggested tags which are provided to effectively reduce common mistakes such as typography errors, and promote tag reuse which will result in a compact tagging vocabulary space.

The second outcome of the research is to provide an improved news system which integrates the controlled vocabulary dictionary with existing system to provide additional functionalities that will enhance current online news systems. Using the proposed system, users can browse news article by utilising the reading tools and can quickly understand the content from the entity in the articles. Based on the controlled vocabulary, tagging functionality will encourage readers to manage and share their articles easily. The structured tagging paradigm is expected to attract users to actively participate in the online news community, especially those who usually have problems in choosing the appropriate tags. In addition, readers who have specific requirements in news consumption can organise their news more efficiently using a tag-based clipbook construction function. In a social context, users will be encouraged to collect and share news articles easily and discover interesting news articles and learn about friends’ activities within the community. This will showcase how a structured tagging functionality can effectively enhance the current news browsing system by providing users with another dimension or facility to organise information. The system extends existing news website by providing additional new features to facilitate the reading process, and aim to provide enhance reader’s reading experience with the least amount of changes in intervening or adjusting the user habits in reading news or tagging.
1.5 THESIS OUTLINE

Chapter 2 provides an overview of online news as the primary case study in this research as it poses a unique type of web publication as compared to wikis and blogs. The main aim of the chapter is to investigate various available systems and the functionality provided for the readers in consuming and aggregating news articles. This chapter also explores the related work in named entity indexing, word sense disambiguation and the choice of knowledge bases currently available. The last part of the chapter presents a discussion of the current online news website limitations which are closely related to the research motivation.

In Chapter 3, an extensive discussion of the current achievements in the folksonomy research is provided. The discussion investigates the type of tagging system that is currently available, tagging practices, tagging motivations, benefits and limitations of current folksonomy systems and lightweight and heavyweight tag structuring methods. The discussion leads to the exploration of user behaviour and folksonomy system features. Lastly, the tag visualisation method is investigated to understand various techniques available for displaying the tags. This chapter aims to provide readers with a wide coverage of the state-of-the-art system which leads to a good understanding of the purpose of this research project.

Chapter 4 presents a high level architectural framework for the entire system. Key components and their functionality are described to provide readers with a big picture for the rest of the thesis which discusses each of the components in details.

Chapter 5 describes the tag indexing and structuring algorithm. The proposed algorithm adopts a lightweight and dynamic approach for use to explicitly define the semantic meaning of a tag and identify the corresponding node in the knowledge base. The mapping model is tested against a large dataset with various types of news articles and is aimed at proving the robustness and the reliability of the model. A dictionary will then be constructed based on the results and output in an industry standard which can then be utilised for the data exchange task.

Chapter 6 describes an online news system which implements the dictionary as a controlled vocabulary for use to bridge the gaps between social tagging and structured annotation. The system implements social tagging and provides features such as reading tools, news clipping and support for named entity based retrieval. The additional functionality enable users to perform various retrieving, organising and information sharing task but without altering much of their existing browsing and tagging behaviour. Various user scenarios are also discussed to illustrate the requirements of the users echoing the current limitation. The implementation details and the new features will be described to explain how the system can effectively overcome the current limitations.
Chapter 7 presents both the experimental results of this research and a preliminary users study. Chapter 8 concludes the thesis and discusses the directions of possible future work and areas of improvement.
Chapter 2: Online News

This chapter provides a review of the current online news systems and works related to entity extraction in the context of this research project. The review focuses on understanding the state of current online news systems, delivery options and limitations. The main aim of this chapter is to describe online news, which is used as the main research domain of this project. This chapter will also discuss state-of-the-art named entity sense disambiguation techniques which are strongly related to the research project, as named entity is considered as the core of a news article. Section 2.1 provides an overview of the trend of current online news systems and Section 2.2 introduces the type of tags that can be extracted from a digital news article. Section 2.3 reports the news characteristics based on 1,000 articles of different genres collected from the source and Section 2.4 discusses the user tagging patterns in an online news environment. Section 2.5 reviews the major news aggregator approach and related personalised news system. Section 2.6 reports the related work of named entity indexing. Sections 2.7 and 2.8 discuss the approach and available knowledge base for use in the word sense disambiguation task. Section 2.9 presents an analysis report of 10 selected online news systems and the chapter is concluded in Section 2.10.

2.1 INTRODUCTION

The rapid growth of the Internet has changed the media landscape development within the past few years and online news has evolved as one of the major channels for the public to obtain information [17]. A survey study from PEW survey[18] shows that about 40% of broadband users consume news from the Internet more than local newspapers and reveals that reading online news has become one of the major ways for the public to learn about breaking news stories and their development. The recent survey by Nielsen Netrating [19] also shows that browsing online news websites is the third most popular online activity (ranked after email and search engine websites). The top 10 parent online companies in Australia survey shows that NewsCorp online (ranked 3rd) and FairFax digital (ranked 8th) outranked other major websites such as Telstra, Yahoo!, eBay and Wikipedia. This indicates that online news is gaining popularity tremendously at a rate that far exceeds any prior estimation. Compared with its traditional publication, digital newspapers provide more instantaneous information about what is happening worldwide, with rich multimedia visual cues and easier accessibility [2]. The advancement in wireless technology and the significant price drops in mobile phone plan access rates has further accelerated the development of such services [20] and hence attracted more
users to consume online news feeds using mobile devices such as PDAs, laptops and 3G mobile phones [21].

There are an increasing number of news providers delivering news content using websites [20]. Major news providers such as NewsCorp Online [22], ABC News [23] and Fairfax Digital [24], publish online editions of news and allow users to subscribe to Rich Site Summary (RSS) feeds to receive direct updates [25]. Their current news presentation model replicates printed copy newspapers by classifying news into categories such as World news, Entertainment, Technology, Sports, Economics, Health and Science. The news is classified by the news creator and presented in reverse chronological order based on the fact that latest news is always of greater interest to the majority and older news is less important. News websites also utilise document indexing, or integrate third party search engines to enable users to perform a keyword retrieval.

2.2 NEWS TAGS

News tags represent the metadata information that can be extracted from the news articles. This information can either be provided manually by the editors during publication time or automatically extracted using annotation tools. In this research, we classify the tags that can be extracted from the news articles into three main categories, namely; creator tags, user tags and named entity tags.

Creator tags represent the descriptive label assigned by the news editor during production time. It is provided by the majority of online news publishers and embedded as keywords in the META tag of the HTML document. Creator tags are mainly used for cataloguing purposes and are seldom made visible to the general readers. ABC [23] and CNET News [26] provide creator tags for users to filter news articles using tags as a subtopic. Creator tags follow specific naming conventions that vary according to different publishers and are well formed, with few spelling mistakes or casual writing.

User tags are short keywords assigned by users to represent their personal perspective towards the content of the article. Social news websites (e.g. Newsvine) allow users to utilise tags to describe the content of the article or function as bookmarks for later retrieval. User tags are not controlled or moderated by any authority, therefore typography mistakes, inconsistent spelling and casual notations are regularly identified.

News articles can be automatically annotated using publicly available software such as GATE [27] and Calais [28] to generate named entity (NE) tags. The tags are automatically extracted and processed using Natural Language Processing (NLP) techniques and can be
classified into appropriate categories according to their nature. The NE tags list provides an overview of key entities in the articles that facilitate an understanding of the content.

2.3 NEWS CHARACTERISTICS

A news article is a type of written publication that reports on a topic of current affairs [29] and is formally written using proper grammar, correct punctuation and capitalisation [30]. News articles utilise named entities to represent four kinds of information namely: who (person), where (location), when (time) and what (keyword) [29]. The story is constructed from the action and relationship between the entities. News articles are ephemeral by nature, the definition and the semantics of the entities are bounded for a period of time and will evolve from time to time. For example, Kevin Rudd was the opposition leader during September 2007 and became Prime Minister of Australia after a month.

The named entity is the key element in any news article. The entity distribution varies according to the nature of the news article. By observing the pattern of the entity distribution, one can understand the importance of each different entity type and their impact towards different types of news genre. Figure 2.1 shows the graphical visualisation of entity distribution from a subset of 1,000 news articles collected from ABC and Newsvine in September 2007. The following observations are made:-

- World, Local, Science and Odd news articles represent location-oriented news where the entities are biased towards the location entity, and indicate the place where an incident or event has taken place. World news and Science news report incidents that have taken place on a national level. In local news and odd news, a person’s name seldom appears as the public’s knowledge of the person being discussed is generally lower.

- Entertainment and political news consist mainly of personal information related to celebrities and politicians. Entertainment news reports on the lives of celebrities, media releases and comments from the journalist. Political news reports on political management, speech rallies, and debates on policy made by politicians.

- Business articles contain mostly organisation information such as public announcements made by a company or the share market or stock index information.

- Event entities generally occur less frequently in news collections.
Figure 2.1. Visualised news characteristics
2.4 USER TAGGING IN SOCIAL NEWS

Named entities play an important role in online news articles. Therefore, it is crucial to understand user tagging behaviour and identify patterns which reveal the type of information with which users are concerned. Unlike other social tagging systems, the tagging vocabulary in digital news is evolving rapidly and the trend of tag usage is difficult to observe. Emerging patterns based on the tags nature can be identified from the tagging space.

To begin with, we present a users tagging activity study to understand the habits of users who perform news tagging to understand their behaviour in correlation to the named entity tags. Figure 2.2 shows the distribution of the user tags from the Newsvine dataset. The dataset consists of 5000 tagging instances from randomly selected users collected from Newsvine data during August 2008. The analysis of this data enable the understanding the tagging pattern from the users while reading online news. The figure reveals that users are taking initiative in assigning tags related to named entities. As we expect, the tags usage in online news tagging are highly correlated to rich named entity feature of news articles. The majority of the tags are used to tag an entity such as a person (23.67%), location (20.00%) or organisation (18.33%). About 11% of the tags are used to describe other entities such as technological terms (e.g. Windows, GPS, Google), medicinal terms (e.g. herpes, glioma, cancer) and products (e.g. wall-e, toyota-corolla, Android). The 14% general terms refer to the usage of non named entity terms which mostly describe the nature of the news closely related to the section such as entertainment, odd and health. For articles without outstanding named entity detected, such as criminal news and general information news, users will apply generic tags such as rape, murder or robbery. A low occurrence of event tags is likely to be caused by the low appearance of events to tag in a news article. In addition, the statistical results from the data set reveal that 78% of the tags assigned by the user are actually named entities (NE) that appeared in the news. This implies that NE tags can be a very useful indicators when guiding the users during their tagging process.

![Figure 2.2. User tags distribution](image-url)
2.5 NEWS AGGREGATOR

This section provides a review on three different news delivery services and their related personalised news delivery systems, namely; Google News (news aggregator), Digg (social news aggregator), Newsvine (community-driven news website).

2.5.1 Google News

Google news is a commercial news aggregator service which provides news stories automatically gathered from various news sources around the world. Google news [31] provides free news aggregation services and presents the news in eight basic categories ordered by relevancy. Users can perform basic customisation functionality to retrieve news according to the source, location or section. Users can also utilise keyword combinations to create their personal scrapbook and to follow up on a specific topic. RSS feeds and email alert services are provided to deliver customised results to the users.

2.5.2 Digg.com

Digg [32] is a social aggregator that receives story submissions from the public and solely relies on users’ participation to moderate the news stories stream. The choice of articles and its popularity are decided by the community and not a professional editorial board. Digg allows users to share and participate in the voting process. A newly submitted story will be placed in the upcoming queue and waits for users to assign Diggs (votes) to it. Users can choose to dig (assign a positive vote) or bury (assign a negative vote) to indicate their personal preference towards the item. Items that receive sufficient votes will be promoted to the front page and items that are buried by the majority will be hidden. As a socially driven website, Digg allows users to mark other users as friends and provides a “Friends interface” for tracking friends’ activities and discovering interesting items. A similar approach can also be seen on websites such as Reddit [33] and DNZone [34].

2.5.3 Newsvine

Newsvine is a community driven social website where the news stream comes from both professional sources (e.g. Associated Press) and amateur writers. Registered users are given a personal column which can be used to seed (post) interesting articles from external sites. Users can also publish news articles written by themselves. Newsvine allows content publisher to assign tags to the article for future retrieval. Other websites that provide similar services include Slashdot [9], a technical news oriented website and OhMyNews International [35], a Korean based social news service. To the best of our knowledge, Newsvine is by far the more popular news system that provides social tagging functionality but it is only limited to content submitter which potentially limits the collaborativeness of tagging.
2.5.4 Personalised news delivery

The needs of automatic personalised delivery rise as the amount of information delivered increases in volume. A personalised news system aims to promote relevant news articles to users based on a model built from the users' preferences. These systems can recommend or automatically rank stories to provide easier navigation based on the content relevancy. Various methods have been proposed to facilitate such functions. In [36] and PNS [37], the system allows users to select topics of interest from predefined categories to create a profile. News articles are then filtered and delivered based on profile modelling results. Users' profiles can also be generated using a data mining technique based on the online news access logs. These approaches are effective when the activity is less dynamic but might not be able to adapt to breaking stories until the profile is updated.

2.6 NAMED ENTITY INDEXING

Named Entity Recognition (NER) is an active research topic in the Information Extraction (IE) field which attracts attention from researchers. It is an important process in the information extraction task that classifies atomic elements in texts into predefined categories such as persons, locations, organisations, time, money, etc [15]. A common application of NER is to annotate unstructured text. For example, considering the sentence “Labor Party leader Kevin Rudd was elected as the 26th Prime Minister of Australia”, it can be annotated as

```
<ENAMEX type='ORGANIZATION'>Labor Party</ENAMEX>leader <ENAMEX type='PERSON'>Kevin Rudd</ENAMEX>was elected as the 26th Prime Minister of <ENAMEX type='LOCATION'>Australia</ENAMEX>
```

NER technique has achieved a great level of success in the past decades and such an application can be witnessed across various Asian languages such as Chinese [38] and Indonesian [39], and in European languages such as German [40], Spanish, Portuguese [41] and Swedish [42] and being applied in news article summarisation [43]. More related to our research direction, it can be used to index documents and provide additional semantics to complement keyword based retrieval. Mihalcea et al. [44] presents an approach to index the named entities in the articles. The system utilised a Minipar tagger [45] to annotate nine types of tags in articles, namely; person, organisation, location, date, number, money, percentage, product and special collocation. The tagging result can then be utilised to answer queries such as “Where is the Taj Mahal?” or “What is the tallest building in Japan?” Users can specify the tag type in query and the returned type. The study conducted an experiment using 125,000 documents from the L.A. Times. The results show that such an indexing method improves the retrieval performance by reducing the number of retrieved documents while still retrieving the relevant documents and that the precision of retrieval is significantly improved (increased by 73%).
Mark Wasson from LEXIS-NEXIS presents a study which shows that named entity indexing can utilise controlled vocabulary terms to incorporate topical queries for news articles and summarising documents [14]. Wasson’s approach extends previous Term-based Topic Identification (TTI) techniques to provide named entity support on company names together with persons, organisations and places with 92% recall and 96.5% precision.

2.7 WORD SENSE DISAMBIGUATION

As the NE index technology is maturing and growing in popularity, another research problem that comes along is the method to locate the appropriate node of an NE term in a dictionary. Word sense disambiguation (WSD) represents a state-of-the-art solution. In folksonomy studies, most of the research focuses on utilising co-occurrence information as context based semantic information as NE are either absent or difficult to identify (e.g. social bookmark, photos, videos). However, in a textual based platform (i.e. blogs, wikis, news) such information is relatively easier to extract. Therefore, to detect tags relationships within the context and to identify the word sense, becomes an important task in structuring the tags. In the past decades, WSD has always been formulated as a supervised learning task [46-49]. Supervised training and a machine learning approach produce a high level of performance and accuracy, however their applicability is limited to the amount of available labelled data, which is strongly connected to the accuracy of sense-tagged data.

As a general approach, Yeung et al. uses a network analysis technique to illustrate that clustering tagged documents can help to understand the semantics of the tags [50]. For example, it differentiates two meanings; “san francisco” and “science fiction,” from the same tag “sf”. The study shows that users are consistent in performing tagging and therefore the sense of the tags is consistent across various different meanings of the tag. This approach does not require any utilisation of a knowledge base which limits the possibility of data enrichment and exchange with other platforms.

In a knowledge-based approach, a number of research studies utilise WordNet to perform Word Sense Disambiguation (WSD). Rattanawongchaiya et al. introduced a three steps model to determine noun phrases in the input context [51]. The process begins by gathering linguistic information of the input term from WordNet (i.e. synonym, definition, examples and hyponym) and attempts to identify the sense by comparing the number of words between the input term and the noun phrase in the given context. If the word sense cannot be identified, the second step will be conducted to predict the word sense using frequencies of synsets and synonyms. If the sense is still undetermined, a sliding window technique will be applied to the top 20 documents retrieved from Google using the term for a similarity calculation and the meaning of the term with the highest similarity is chosen as the sense. Similarly, Canas et al.
proposed a six step process to utilise WordNet information to identify the sense of a term in a concept map [52]. The algorithm begins by selecting key concept candidates to be included in WSD, the terms then undergo morphological transformations and identify the senses within WordNet. The synsets are then grouped into clusters using their hypernym distance, based on a single synset per term in each cluster. The synset of the cluster with the highest weight containing the term will be chosen as the sense.

As the information grows, researchers have come to realise that the update frequency of WordNet is relatively slower and unable to cover most of the named entity topics such as politician, celebrities and events. Therefore, researchers turn to an alternative knowledge base to provide better support on named entities. Nguyen and Cao proposed to use KIM ontology to perform knowledge-based and class-based WSD for named entities [53]. In knowledge-based ranking, co-occurrence information is utilised to perform pair wise disambiguation. For example, if the term Brisbane is co-occurring with Queensland, it will be decided that it is referring to the Brisbane, Queensland (Australia) instead of a city in California, America. In class-based ranking, preferences are given to preferred class based on the assumption that common classes are more likely to appear than others, e.g. “Athens” is more likely to be the capital of Greece rather than a small town in Georgia, United States. Lastly, a shallow co-reference method is applied to identify possible entities in the same text chain (i.e. to detect a main alias for a full name). For instance, if “George W. Bush” appears in the article, subsequent appearances of “George Bush” and “Bush” will be co-references to “George W. Bush” instead of “Barbara Bush” or “Laura Bush”.

On the other hand, as the popularity of Wikipedia increased, researchers investigate the possibility of utilising it as a knowledge base in WSD to overcome various bottlenecks in the previous methods. The following key properties of Wikipedia are noted:-

- **Unique identifier:** Each article in Wikipedia is uniquely described by a fully qualified identifier to ensure each wiki page is concisely describing a singular topic with no ambiguity. For example, Mercury can be uniquely identified as Mercury (planet), Mercury (element) and Mercury (mythology).

- **Disambiguation page:** Wikipedia provides a unique feature to gather the possible senses of an ambiguous term. For example, a disambiguation page for George Bush shows the possible senses of George W. Bush (current U.S. President), George H. W. Bush (41st U.S. President) and George Bush (NASCAR), a former NASCAR racing driver.
• **Redirect pointer:** For semantically identical wiki pages, Wikipedia provides a redirect hyperlink in the wiki page instead of the full article to indicate the synonymy. For example, Peking is automatically redirected to *Beijing* upon request.

Mihalecea exploits the above mentioned Wikipedia structure to build a sense tagged corpora for WSD [54]. The text is pre-processed by tokenisation and part-of-speech annotation. A sliding window is then applied to form a compound concept defined in WordNet. A local feature set is formed by selecting *current word, three words from left and right of the ambiguous word, part-of-speech of surrounding words and verb and noun before and after the ambiguous word*. Five words that appear at least three times in the context are selected as the global feature. A Naive Bayes classifier is then constructed to perform word sense disambiguation.

Similarly, Bunescu and Pasca utilise the Wikipedia facility to build an SVM taxonomy kernel using a cosine similarity based model [55]. This system utilises an optimised window using 55-word contexts around the entity to perform sense detection and disambiguation across various datasets and achieve accuracy between 55.4% to 84.5%. A similar approach to our work will be Cucerzan which performed a WSD using entities from the document and Wikipedia [56]. The study analyses Wikipedia dump of September 2006 (about 1.4 million entities) and reports average 2.4 surface forms per entity. Lexico-syntactic patterns such as titles and cross referenced links are used to build the co-references in the context. Feature vectors are constructed based on documents and wiki pages and the wiki page with the highest similarity will be considered as the sense of the entity in the document. This model achieves an 88.3% accuracy using 350 Wikipedia articles and a 91.4% accuracy using 100 news articles from MSNBC news.

These approaches show that Wikipedia is suitable to be used as a knowledge base in a WSD task for its rich amount of information and regular information reviews. Wikipedia will be adopted in this research as the utilised knowledge base, however a different approach without training and predefined window context will be applied.

### 2.8 KNOWLEDGE BASE

A main function of our indexing approach is to map the tags to the knowledge base (KB). The performance of a mapping algorithm depends on its capability of identifying concise meanings to the KB and the quality of a KB relies on its ability to cover more topics and group identical terms together and providing clear meaning distinctions for each ambiguous term which shares the same form. In this section we will discuss three main knowledge bases that have been used in the literature, namely; KIMO ontology, WordNet and Wikipedia. Knowledge and Information Management Ontology (KIMO) [57] is a semantic platform developed by OntoText which provides infrastructure and services for automatic semantic annotation. KIMO allows
upper-level resources classification using 250 entity classes and 100 attributes and relations with the data source populated by NE, extracted using GATE [58]. WordNet is a semantic lexicon constructed for the English language which groups English word into sets of synonyms named Synsets. A semantic relationship between each individual synset is defined to provide a combined dictionary and thesaurus to support automatic text analysis, for example, Laniado et al. [59] employed WordNet to construct a hierarchy of folksonomy tags, however it is reported that important terms such as “JavaScript” and informal abbreviations such as “nyc” are not recognised and external domain ontologies such as Wikipedia are required to address these limitations.

Wikipedia [60] is a collaborative knowledge base constructed voluntarily by public contributors in a collaborative writing manner. It provides a wide range of named entity topics covering various domains including information technology, biology, business, sports, technology, people, facilities, etc. Wikipedia content is dynamic and growing quickly where articles about news events generally appear within a day of the occurrence. The articles count increased from 751,666 articles to 1.4 million articles from September 2005 to September 2006 and to date, English Wikipedia alone already consists of approximately 2.5 million articles. In terms of its applicability in handling various common linguistic problems, Wikipedia also provides a facility for use to identify synonyms and resolve ambiguity using a redirect disambiguation page. Redirection (Table 2.1) is used to model various synonymy situations (such as spellings variations, abbreviations, and alternative names) therefore different surface forms of an NE will only be pointing towards a singular main root form. The disambiguation page (Table 2.2) is constructed to denote every possible meaning of an ambiguous NE to show the distinctions among different meanings. Swift evolvement in Wikipedia contents provides greater advantage over static ontology as it is able to capture more updated information about current topics such as technology, celebrities and events which are generally not available in ontology.

<table>
<thead>
<tr>
<th>Wiki page title</th>
<th>Redirect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUT</td>
<td>Queensland University of Technology</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>Beijing Olympics</td>
<td>2008 Summer Olympics</td>
<td>Administrative redirect</td>
</tr>
<tr>
<td>Tae Kwon Do</td>
<td>Taekwondo</td>
<td>Spelling variation</td>
</tr>
<tr>
<td>Suu Kyi</td>
<td>Aung San Suu Kyi</td>
<td>Alternative name</td>
</tr>
<tr>
<td>Rangoon</td>
<td>Yangon</td>
<td>Alternative name</td>
</tr>
</tbody>
</table>

Table 2.1. Examples of Wikipedia redirection
Table 2.2. Wikipedia disambiguation of “NBA”

<table>
<thead>
<tr>
<th>Wiki page title</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>National BasketBall Association</td>
<td>American national basketball league</td>
</tr>
<tr>
<td>National Boxing Association</td>
<td>Former name of World Boxing Association</td>
</tr>
<tr>
<td>National Braille Association</td>
<td>Organization for vision impaired education</td>
</tr>
<tr>
<td>3-Nitrobenzanthrone</td>
<td>One of the chemical in diesel fumes</td>
</tr>
</tbody>
</table>

2.9 LIMITATIONS OF CURRENT SYSTEMS

While users can gain access to information easily and receive information from multiple sources with the aid of news aggregator, several areas in the current online news model still need to be improved in order to provide users with features that help to gain control over the information presentation. This section examines 10 online news websites listed in Table 2.3 and compare their functionality from 5 key perspectives: reading, retrieving, organising, managing and sharing.

Table 2.3. Online news website used in the study

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Australian Broadcasting Corporation</td>
<td><a href="http://www.abc.net.au">www.abc.net.au</a></td>
</tr>
<tr>
<td>2</td>
<td>The Sun</td>
<td><a href="http://www.thesun.co.uk">www.thesun.co.uk</a></td>
</tr>
<tr>
<td>3</td>
<td>Cable News Network</td>
<td><a href="http://www.cnn.com">www.cnn.com</a></td>
</tr>
<tr>
<td>4</td>
<td>The Sydney Morning Herald</td>
<td><a href="http://www.smh.com.au">www.smh.com.au</a></td>
</tr>
<tr>
<td>6</td>
<td>Reuters</td>
<td><a href="http://www.reuters.com">www.reuters.com</a></td>
</tr>
<tr>
<td>7</td>
<td>The Associated Press</td>
<td><a href="http://www.ap.org">www.ap.org</a></td>
</tr>
<tr>
<td>8</td>
<td>Wall Street Journal</td>
<td><a href="http://www.wsj.com">www.wsj.com</a></td>
</tr>
<tr>
<td>9</td>
<td>Newsvine</td>
<td><a href="http://www.newsvine.com">www.newsvine.com</a></td>
</tr>
<tr>
<td>10</td>
<td>The Australian</td>
<td><a href="http://www.theaustralian.news.com.au">www.theaustralian.news.com.au</a></td>
</tr>
</tbody>
</table>

2.9.1 Lack of reading time support

Reading tools that provide browsing time support for users are currently not available in any of the online news websites. At the moment only related articles are presented by news
where websites should provide an interface that enables users to know more about the background information [61]. The news presentation begins with a large heading using a font weight and colour chosen to stand out, followed by a paragraph that summarises the news content. Users can choose related articles at the end of reading the whole article. This reading mode requires a user to read through the whole article in order to understand the content of the article which can cause the user to spend a certain amount of reading time depending on the length of the article. Information such as entities, information about entities and related entities are not provided. Users can only rely on the search function of the web browser to manually locate interesting terms which is inefficient and might not be accurate all the time. Searches need to be done repetitively if the users are interested in multiple terms.

2.9.2 Support for Named Entity based Search

While online news systems are the main channel for obtaining real time information, a common problem is that users are often overloaded with too much information. To effectively locate and retrieve information from the news streams remains a challenge. This phenomenon becomes more apparent when using aggregator system such as Google News and Digg which consistently aggregates news article from various sources.

The majority of news websites currently only provide users with basic keyword-based search functionality using traditional document indexing, or delegate the task to an external search engine such as Google. Advanced search functionality is occasionally provided, however it is still limited to filtering articles according to the publication time or the location of the keyword (e.g. title, section or content). Some websites also enable users to utilise Boolean query operator (‘+’, ‘-’, ‘AND’, ‘OR’) to explicitly include and exclude a term, but this functionality is seldom promoted. Due to the absence of controlled vocabulary, users are suffering from either obtaining less results by using a strict operator (e.g. exact phrase “Paris Hilton”), or humongous result using a joint operator (e.g. “Paris” or “Hilton”). There is no clear indication of the returned result quality and users will need to perform multiple iterations to experiment with various keyword combinations in order to achieve an optimised result. In addition, the search terms and search result usually cannot be stored for future retrieval purposes.

2.9.3 Organisation tools

Resource management and organisation is one of the main challenges persisting in information retrieval. As the amount of articles increases, an organisation mechanism is required to assist users in managing the content effectively. The majority of news websites do not support users in performing any onsite content organisation at the present moment. Users either rely on the bookmark function in the web browser or submit the links to an external bookmarking
system. Article management and relevant activity such as tagging is then performed at the external site. For example, a user reading an interesting article on CNN and wishing to bookmark the article will have to copy the URL to Newsvine and then assign tags to label the article which can be used for future retrieval. This process allows users to organise contents according to their preference but the efficiency declines when the amount of articles is large and the activity is frequent. This causes the users to switch between websites and potentially increase the burden on the users for organising bookmarks.

2.9.4 Content management

News value depreciates rapidly and headlines evolve swiftly over time, which leads to difficulty for readers in following up breaking news stories of an event or topic of interest. Topics of interest in news websites are decided by the news editors according to their professional judgement and by the trend of the majority of public. Users are unable to form topics based on their personal preferences and always face difficulty in retrieving backdated news articles. News articles can be easily missed out and unfindable if it is no longer visible in the headlines. News reader services such as Google Reader [62] and Bloglines [63] are designed to solve this problem by helping users to collect the RSS feeds periodically for users to read at a later time but there is no mechanism available that helps the user to filter the content according to their needs. To effectively browse through huge collections of news stories is a crucial requirement needed to help users in selecting their preferred news articles more quickly. In addition, headlines that have been read will be archived automatically unless specifically marked. This imposes a certain level of difficulty for users to keep track of their reading history.

An alternative method that can be used for learning about breaking news is by utilising news aggregator services (e.g. Google News). News aggregator collects news articles from various sources and utilises text mining and clustering algorithms to group similar articles together and present them as a topic. These functionalities obtain information automatically and provide users with related reporting of different perspectives from different sources. Users can easily retrieve news articles using terms such as “Kevin Rudd”, “Japan” to retrieve news about Australian Prime Minister Kevin Rudd visit to Japan, however, the limitations remain when trying to answer complex queries such as “How is the Australian Financial Market in the past 3 days?”, “Who has been related to Kevin Rudd lately?” and “How does the drop in Nikkei index affect Australian market?”.

2.9.5 Scattered social activity

The development of folksonomy in Web 2.0 has caused the advancement of the social website which facilitates users in communicating and sharing interesting information among the
community. News websites begin to include hyperlinks to social systems which provide users with an additional channel to share news with friends over and above the sending of email. Despite the current availability of social bookmarking websites (e.g. Delicious, Facebook) which facilitate item posting using an immediate sharing mechanism, users are still unable to track or capture the general overview. Therefore, it is more viable and sensible for an online news provider to exploit the simple and easy nature of tagging and to implement the tagging service directly on the website rather than relying on a third party social bookmarking service provider. By doing so, users gradually changed from sending hyperlinks to friends using email to share in the social websites. This approach successfully facilitates social information sharing, however the problem in such a model is that users are required to alternate between different websites to find information.

2.10 CONCLUSION

This chapter presents an overview of the currently available online news systems from the perspective of their functionality and features. Online news systems have indubitably become an important part of web use, as a major channel of receiving updates and learning breaking stories. The content delivery can be easily accessed via RSS feeds and personalised using available algorithms and collected automatically using feeds readers or through news aggregator services.

While the news content is easily accessible compared to decades ago, users are now looking forward to more functionality to gain control over the presentation. From document indexing and content analysis perspective, TF-IDF remains the dominant technique together with using cosine similarity as a metric. Term based searches remain the main method of retrieving information, but the limitation comes from the users only being allowed to perform queries using various keyword combinations. Additional semantics through a different indexing scheme such as NE based indexing are still required in order to provide better support for complex queries. The organisation and sharing functionality provided are still in a simple stage and most of the tasks cannot be performed onsite, where users are required to perform multiple registration and familiarisation with different tools even to perform basic bookmarking and sharing.

In the next chapter, we present the study of folksonomy, a community driven based activity which allows users to manage and share information collaboratively with the public. The advantages and disadvantages are investigated to evaluate the appropriate techniques for tags structuring.
Chapter 3: Folksonomy

This chapter presents an insightful discussion of related work in the folksonomy study. Section 3.1 introduces the key concept of folksonomy and its main applications. Section 3.2 and 3.3 describe various tagging practices and the motivation behind using them. Section 3.4 and 3.5 detail the different types of folksonomy and tagging systems. Section 3.6 and 3.7 provide an extensive review of the advantages and disadvantages for using social tagging. Section 3.8 reports the main work on tags structuring and Section 3.9 reviews the study of tagging users. Section 3.10 evaluates the available tags visualisation approach and Section 3.11 summarises the features of social tagging. Section 3.12 then concludes this chapter.

3.1 INTRODUCTION

The development of the social websites has been advancing and gaining popularity at a tremendous rate. It presents one of the key activities addressed in Web 2.0 [3]. In addition to content sharing, social websites such as Delicious, Flickr and Youtube implement free form tagging functionality which allows users to assign freely chosen keywords as tags for content organisation and social classification. Van der wal first coined the term ‘folksonomy’ (a combination of folks and taxonomy) to describe the process of generating taxonomy from the user tags. Social activities such as social bookmarking and collaborative tagging have drawn significant attention from both industry and academia in the recent past [64, 65].

Folksonomy services have experienced rising popularity in the past few years. PEW Internet survey of December 2006 [18] reveals that 28% of American internet users have utilised a tagging service for marking online content such as photos, news or blog posts and the amount of users is currently increasing at a impressive rate. At the moment, users can submit a URL as a bookmark using Delicious [7], organise personal photography using Flickr [5] and share video using YouTube [66]. Such options are also recently being implemented on social news websites such as Newsvine [8] and Mixx [67].

The main idea behind social tagging activity is to avoid any formally defined hierarchy as the backbone and provide every user with absolute freedom to annotate the resources. Users are encouraged to utilise tags to reveal their interpretation of the resource and this process is known as tagging. This phenomenon caused social tagging to become the leading activity in community-based content sharing websites and draws attention from academic researchers to conduct studies on websites such as social bookmarking sites, wikis and blogs.
The tagging process is not a brand new concept, as similar techniques have been broadly applied in the past during different stages of information retrieval such as manual indexing, data training and ground truth building. In social tagging websites, users are required to register or create an account through a setup process and they can begin to upload and share content after login. Upon completion of the upload, users will be prompted to provide tags for describing the uploaded content using a choice of keywords. The tags represent the users’ perspective of the content which can be used for self-retrieval in the future, or to retrieve other related items under the same tag.

As the amount of tags increases, the user may start to find it difficult and time consuming to browse through all the tags and needs to visualise related tags and popular tags. Therefore a tag cloud arises and becomes the powerful visualisation to address the users need. In short, a free form tagging system is a powerful manual classification technique which requires little computational power. With the collaborative effort, the system is able to categorise and group related items together quickly and efficiently, to identify the relationship among user groups and their common interests, which can be utilised as an information retrieval tool with the ability of social recommendation.

3.2 TAGGING PRACTICES

The idea of tagging, or the similar approach of assigning multiple topics to a singular document has already been widely adopted in the past decades. For example, in 1996, the HTML 2.0 specification already allowed webpage creators to provide multiple metadata to a document by embedding keywords into the HTML “META” tag [10]. This information is used by the search engine for indexing and is seldom made visible to the user's. Recently, tagging functionality has been widely implemented in various different social and personal systems and the tagging practices can be classified into the following according to the activity nature:-

3.2.1 Public tagging

Public tagging allows any users to tag concurrently on the same resources using their own tag with no restrictions or limitations. This practice can be seen in the social tagging system where users upload bookmarks and assign tags collaboratively. Users who participate in public tagging will have the right to influence the tagging space. The key characteristic of a public tagging model is that the output will be dynamic and the tagging usage will directly reflect the public’s point of view on the specific resource.

3.2.2 Creator tagging

Creator tagging stipulates that only the content submitter, creator or owner are allowed to annotate and change the tag of a resource. This tagging practice is adopted by content publishing
based websites such as news, blogs and wiki. A typical example of such will be NEWS.COM [68] and ABC News [23] which assign tags on published articles to show the nature of the content and subtopic of the news. Blogging software such as BlogSpot [69] and WordPress [70] implement the tagging feature as part of the core function of the system to encourage bloggers to tag their blog post, which can then be used for classification and assist readers in retrieving relevant blog posts.

### 3.2.3 Machine tagging

Apart from the two manual tagging methods mentioned above, machine tagging (also known as automatic tagging) is another frequently used technique in generating tags for articles. Machine tagging applies the TF-IDF (Term Frequency Inverse Document Frequency) [71] technique to calculate the weight of every term in the document and identify the top ranked terms to be used as the basic tags set. This set of tags shows the importance of the keywords in the article with respect to the whole collection. Brooks et al. reports the result of a research study using 500 automatically tagged articles using TF-IDF score [72]. The result indicates that machine tags are suitable for use as automatic tags for articles. The top three terms with the highest TF-IDF score are selected for clustering. The result reveals that the automatic tag produced more focused topic clusters while compared with human tagging, this shows that automatic annotation techniques can potentially benefit users who intend to search for articles based on a specific subjects.

### 3.3 TAGGING MOTIVATION

The purpose of utilising tagging application changes with online sharing communities depending on the subject area. However, the users share common reasons for such an activity which motivates the users to annotate contents. Traditionally, tagging might be a simple function for users to manage the content in a casual way, in the current context, the functionality of tagging has been extended and can be utilised to become a personal management, self-promotion, sharing and resource discovery tool. A key paper by Golder and Huberman [64] identifies the primary features of a tagging system employed by the users (Table 3.1):-
Chapter 3: Folksonomy

Table 3.1. Golder and Huberman tagging feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Function</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify what it is about</td>
<td>To identify and exchange ideas about people and organisations</td>
<td>Paris Hilton, Federal Government</td>
</tr>
<tr>
<td>Identify what it is</td>
<td>To indicate the nature of the item</td>
<td>Javascript, iphone</td>
</tr>
<tr>
<td>Identify ownership</td>
<td>To identify the content creator</td>
<td>ScottGuthrie, WarrenBuffett</td>
</tr>
<tr>
<td>Refining categories</td>
<td>To qualify existing tags</td>
<td>Playoff2008, worldcup2010</td>
</tr>
<tr>
<td>Identify characteristic</td>
<td>To show tagger’s opinion of the content</td>
<td>Funny, scary, inspirational</td>
</tr>
<tr>
<td>Personal reference</td>
<td>To indicate the relationship between the content and the tagger</td>
<td>Myblog, mycontent</td>
</tr>
<tr>
<td>Task organisation</td>
<td>To group items according to task based information</td>
<td>toread, jobsearch</td>
</tr>
</tbody>
</table>

Kipp and Campbell conducted an insightful study of tagging practices [6]. The paper analyses the tagging pattern by studying the behaviour of users from Delicious. The research aims to understand how collaborative tagging enhances the traditional document classification and indexing scheme. Descriptive frequency data including number of posts per user, number of tags per post, number of unique tags, number of unique tags with specific frequency, number of users with specific frequency and number of users with specific number of tags, are all analysed. Co-word analysis is conducted to examine the correlation of word-pairs that co-occur frequently. For e.g. if tag x and y are used by 3 users together on the same post, it is said that the co-occurrence of x and y is 3. The research shows that up to 65% of delicious users assign 1 to 3 tags to a URL and only 6% of users do not assign any tags to the URLs they bookmarked. The analysis results suggest that several predictable patterns can be anticipated easily and aid in terms suggestion, and synonymy can be detected which helps in maintaining the fundamental dictionary.

Tagging motivations range from personal to public sharing purposes. Hammond et al. presents an insightful study of users tagging motivation [73]. Figure 3.1 illustrates the benefits of tagging on the vertical axis against the content being tagged on the horizontal axis. Quadrant 1 represents tagging practices in which users utilise tags to indicate the published content (e.g. blogs). Quadrant 2 represents users who are annotating others resources for other users (e.g. Wikipedia). Quadrant 3 represents individual users who tag content for personal consumption without considering the general public’s point of views (e.g Flickr, Gmail). Quadrant 4 represents
the phenomenon in which individual users explore and annotate other users' contents for their own use.

Ammes and Naaman conducted a study using ZoneTag and Flickr to understand the motivations behind mobile and online media users. Figure 3.2 shows four dimensions of sociality from the study result. Quadrant 1 indicates that users utilise such applications for organising their own photos to ease the future retrieval process, by using tags to annotate events or persons (e.g. Brother's Wedding, My Children). Quadrant 2 represents users who applied tagging functionality to provide additional context to the images for memory recall purposes (e.g. name of the person in the picture, name of the location where picture is taken). Quadrant 3 represents groups of users who tag the photos for sharing among family and friends or facilitate discovery for public users who may find the pictures interesting. Quadrant 4 represents the group of users who annotate photos to publicise their movements and provide additional descriptive information about their personal interests (e.g. photographer to publish their work, travel journalist to report about their journey). The study results agree with Hammon’s findings and indicate that social bookmarking systems are a tool that benefits users in both personal content organisation and public content sharing.

Figure 3.1. General tagging motivation
3.4 TYPES OF FOLKSONOMY

Vander Wal proposes that folksonomy can be classified into broad folksonomy and narrow folksonomy [74]. Broad folksonomy systems allow single resources to be tagged more than once (e.g. Delicious, Technorati) and multiple users can assign multiple tags on the same resource concurrently. Figure 3.3 illustrates the graphical representation of a broad folksonomy. Content creator submits content to the system and assigns tags, subsequent users who view the contents will also see the tags that were assigned by the creator. This user can utilise and reuse the tags assigned by the creator or other users or assign new tags for their own personal use. Broad folksonomy generates a rich tagging space by capturing information from multiple users. The statistical information produced by the broad folksonomy tag usage will directly reflect the viewpoint towards the content from the general public perspective.

Figure 3.2. Tagging motivation in photo sharing

Figure 3.3. Broad folksonomy ([74])
Narrow folksonomy systems (e.g. Flickr, Youtube) allow tags to be solely assigned by content creator or submitter (Figure 3.4). The tags represent the content meaning from the owners perspective. Content readers can utilise the tags to understand the classification and retrieve relevant items, however they cannot make any changes or suggestions with relation to the tag. This type of system is commonly used for personal content management and publication (such as photo, videos and blogs) where creators are responsible and in control of their personal taxonomy and the public’s opinions are less significant.

The outcome of both folksonomy styles will result in a tag vocabulary space that follows a power curve and shows an effect of a long tail as shown in Figure 3.5. This describes the phenomenon in folksonomy where a small amount of popular tags are used to annotate a large amount of content, while a large amount of infrequently used tags are sparsely spread across a smaller amount of collections. The long tail technically causes the drop in recall rate of a tagging system and a large amount of tags is required in order to efficiently cover more content across different users perspectives.
3.5 TYPES OF TAGGING SYSTEMS

Folksonomy systems can be classified into three main types based on their functionality, namely; social bookmarking, personal tagging and collaborative tagging, as depicted in Figure 3.6.

![Figure 3.5. Long tail of folksonomy](image)

![Figure 3.6. Folksonomy hierarchy](image)

A basic social bookmarking system is an online repository which allows users to store interesting or important URL’s of a website for future retrieval and sharing purposes. Bookmarks are archived at an online repository and not through the browser in the local machine. This allows users to easily access the bookmark at a later time at any place where Internet connection is available, without being limited by their physical location. For example, Furl [75] allows users to collect bookmarks and explore relevant bookmarks of other users. Tagging mechanisms are not included in the design of such a system but users can organise the bookmarks into predefined categories such as sports, entertainment and technology. Users can browse the list of popular bookmarks based on the number of users who added the bookmarks. Implementation of such functionality is recently seen in social media aggregations such as Digg and Reddit. These systems behave as resource collectors and recommenders which allow users to submit and vote.
on interesting stories. The system then recommends content to the users based on the voting outcome.

Personal tagging systems are a typical narrow folksonomy system that only allows the content owner to assign a tag to the content. The system plays the role as a content manager to assist users in managing personalised content, as the public's viewpoints are less important. The authors decide the vocabulary based on their own knowledge to decide on the tag that is best to describe the content. The primary function of tags in these systems is for grouping similar items to facilitate future retrieval. This model resembles a traditional classification process where a cataloguer is responsible for classifying content within a domain using a set of taxonomy but without a predefined control vocabulary. The increased popularity of such trends is noted in recent systems that enable users to organise their own media such as Flickr (Photos), YouTube (Video), Google Mail (Email) and Windows Vista (Files).

Collaborative tagging describes the process of multiple users assigning multiple tags within a resource. This allows users to publicly annotate the content and share it easily with other users. As opposed to personal taxonomy, collaborative tagging systems generate a large community vocabulary which describes single content using different terms from different perspectives. Collaborative tagging systems are useful in promoting social sharing. A good example of such systems is Delicious where users submit bookmarks to the system and allow others users to collaboratively tag on the same content. A handful of social statistics information can be generated from the tagging result to display, e.g. the popularity of content (number of users who submitted a bookmark), users tags (number of tags applied on the bookmark), etc. Tagging space appears chaotic initially but it is proven that the tagging space will be stabilised and result in a tag vocabulary that represents the common language spoken in the community over time [76].

3.6 BENEFITS OF FOLKSONOMY

The rapid growth of social tagging services is highly credited to the following immediate benefits generated for the users:

3.6.1 Low overhead cost

Social tagging requires relatively low cognitive costs when compared to rules based classification. Top down classification systems require a sets of rules and controlled vocabulary be defined prior to performing content classification. Users are required to learn and familiarise themselves with the hierarchy and classification theory before utilising structured annotation. Users without prior knowledge in the classification technique are either not allowed to perform the operation or will possibly misclassify the resources, which decreases the overall information
retrieval performance. In contrast, flat style tagging allows users to categorise resources using keywords that do not need to conform to any predefined rules or hierarchical taxonomy [50]. As proposed by Sinha [77], tagging is a two steps process (Figure 3.7) where taggers only need to input tags for the immediate activated concepts. Categorisation requires user to select one of the activated concepts cautiously which can potentially lead to analysis-paralysis (inability to select the most appropriate concept), which might occur while performing the task (Figure 3.8). The additional concept selection phase incurs additional cognitive costs to user which might cause the user to be hesitant with their involvement in classification. While active arguments exist between social tagging and categorisation and we agree that attracting more user participation will be a worthwhile trade-off for the imprecision caused by social tagging, the problem can be solved by using controlled vocabulary to improve the quality of social tagging efficiently.

Figure 3.7. Cognitive process of tagging [77]

Figure 3.8. Cognitive process of categorisation [77]

3.6.2 Finite result

The tagging process can be considered as a manual indexing and social classification process. As opposed to a keyword-based search where users are unable to tell whether the query has returned all the relevant results, a tagging system ensures users can be certain that the tags will retrieve the entirety of documents collected under the tag [78]. It is expected that a user will be able to retrieve more relevant results if tagging is performed effectively. A tag efficiency study from Brooks and Montanez [72] shows that even though tagging is not very effective in assisting users to determine a specific topic of an article, it was undeniable that tagging generates a lot of
benefits and helps to promote the relationship in the community. This facilitates community information sharing and is effective for grouping articles into categories. The number of recalls is expected to increase and strengthen the weaknesses in folksonomy, which can significantly improve the precision of tags retrieval.

3.6.3 Social context sensitive

The collaborative tagging process is a self-organised social feedback mechanism whereby each tag represents the users knowledge level within a domain and global tagging space, revealing the social context of the community. Tags are directly input by users and the resulting tagging vocabulary plays the role to form the basis of social navigation and shared expression by reflecting the user language and indicating the topics that draw the most user attention. Sen et al. [79] further confirm this argument by showing two important aspects: Personal tendency and community influence which means that users provide tags according to their personal preferences and experience in the knowledge domain and the tags distribution gets stabilised over time as a result of influence among the community members. This shows that users are eligible to create personally meaningful tags, which can certainly be used for information management. Furthermore, in a dynamic environment, new terminology appears frequently and it is usually discarded after a short period, therefore learning the vocabulary is not cost effective and it declines user interest in participation. Free form tagging fills in the gap by giving users absolute freedom to select any keyword to represent the level of understanding among themselves or a group within the community. This helps to save a significant amount of time, by avoiding having to learning the relevant terminology, and eventually attracts the user to participating actively in classification.

3.6.4 Easy implementation

Another free form tagging characteristic is that its implementation is relatively easier and simpler as compared to classification. Folksonomy does not require any pre-processing process to capture all the entities and determine their relationship from the data collection. Classification requires a set of rules to be established initially for use in cataloguing the content. For hierarchical classification to be integrated into the system, the user interface is required to be redesigned to guide users in browsing and editing through the preset hierarchy. Tagging functionality can always be implemented as an add-on feature to existing websites without drastically changing or intervening in the existing interface and remains as an optional function for users.
3.7 SOCIAL TAGGING DISADVANTAGES

While social tagging appears to generate numerous immediate benefits, various problems still exist in folksonomy due to its free form nature. The three main areas that have been identified are listed below.

3.7.1 Low precision

Word ambiguity is an inherent problem that exists naturally in the English language. It is also described as polysemy in the literature to describe the phenomenon of a word with multiple meanings [80]. For instance, the term “Apple” can refer to both the fruit or computer products manufacturer [64,78] and the word “Java” can refer to a programming language or an island in Indonesia. In the case of abbreviation, meanings of a term vary based on the domain of application. For example, the term “KFC” is internationally recognised to represent Kentucky Fried Chicken but it can also refer to Kangaroos Football Club when used in Australian local news. The inadequate support in identifying or disambiguating the word sense within polysemy causes a drop in the precision rate, causing irrelevant information to be retrieved. This eventually restricts the data exchange possibilities with other systems, as accurate word sense cannot be identified.

3.7.2 Low recall

Low recall rates in folksonomy are caused by synonyms. Synonymy refers to different words that have the same meaning or a similar meaning. In a social tagging context, it refers to multiple different tags that are used to annotate resources which are semantically similar. This problem is well documented in various literatures and is regarded as the main issue in the classification. The problem is difficult to be tackled as it is difficult to identify the number of variations of the words to identify in advance [78,81]. This is recognised as the main drawback of social tagging, as it leads to the drop of recall in tagging systems.

One of the commonly seen examples is the use of an organisation name which appears in both abbreviated form and root form such as Federal Bureau of Investigation (FBI, F.B.I), United Nations (UN) and Europe Union (EU). For location entity, some countries and cities have alternative names and some have undergone name changes for political reasons, which causes the location to possess both their original and international names. For example, the capital of China was spelled as ‘Peking’ before the Chinese government adopted Pinyin and is now written as “Beijing.” Locations under special political conditions such as Taiwan are also known as Chinese Taipei and the Republic of China, which further complicates the problem.

Another complex problem is the person entity. Different writing cultures and naming conventions between different sources can affect the surface form. A classic example is the
current United States President George Walker Bush. Commonly it is written as “George Bush” in most contexts or in the abbreviated middle name form “George W Bush” or “George W. Bush.”. Another observation is the variation of the names of celebrities, with their nicknames, stage names and actual names. For instance, American artist Queen Latifah is written as ‘Dana Elaine Owens’ in formal news articles, but her stage name is generally used in entertainment news. Furthermore, naming conventions vary in different cultures. For example, in Chinese culture, a name is spelt in the surname-given name format (Yao Ming) instead of given name-surname format (Ming, Yao). Modern Chinese also uses a two character given name to indicate generations in the family. It should also be noted that removing or abbreviating the second character in the name is inappropriate (i.e. ‘Ze’ in the name ‘Jiang Ze Min’).

3.7.3 Lack of control

Social tagging vocabulary tag space grows unlimitedly without the presence of a control vocabulary. While social tagging activity aims to provide extra ordinary support for serendipity exploration and discovery, the uncontrolled tags in fact increase the difficulty for users in locating specific information.

The absence of control vocabulary and tag suggestions causes the problem of ‘long tail’ in folksonomy. Large numbers of tags have been created by a small number of users with a very low reuse rate. Different support in tagging practice has been introduced in various systems, which further complicates the situation. For instance, delicious and flickr support single word tags only, so that users turn to other alternatives such as using hyphens, underscores and no spacing in tagging which causes the retrieval to be difficult. Experienced tag users can overcome these problems by experimenting with various different combinations (e.g. brinney_spears, Britney spears, Britney-spears) to obtain similar results, but inexperienced taggers might find the process complicated and be confused by the additional steps incurred. This has limited the system capability to exchange information with other systems.

3.8 TAG STRUCTURING

Various approaches have been proposed to overcome the problems in the folksonomy system and tags structuring is always presented as the main solution to solve the problem. The primary goal in tag structuring is to explicitly identify and disambiguate the different meanings of the tags. This can be achieved by monitoring the usage or interpreting the meaning from the derived information from the co-occurrence relationship. A controlled vocabulary dictionary is commonly used to represent the unique meaning of a specific tag.

A widely accepted approach in structuring folksonomy is to apply predefined ontology to manage the social tags. The application of ontology produces highly precise results but the
static structure limits the scalability of folksonomy. In addition, users are required to familiarise themselves with the rules and schema prior to performing tagging. The other approach is to employ data mining techniques using statistical information derived from the tagging space such as vocabulary usage and co-occurrence frequency to understand the underlying emergent semantics from the tags. This approach is more dynamic and aids in identifying topics easily, but regular operations such as clustering and model training must be utilised and additional inputs are required to interpret the result.

Previous studies conducted on social tags structuring are mostly based on general tags usage, which focuses on various types of nouns and relatively lower amounts of research are performed on named entity tagging. As the usage of tagging is increasing and gaining in popularity in dynamic environments such as wikis, blogs and news, users start to shift towards using named entity tags to describe the contents and aid in classification. Rapid evolution in the vocabulary space is not easily traceable and patterns are harder to identify, which becomes another challenge for the previous methods.

In this section, we present various state-of-the-art research studies that have been proposed to structure the folksonomy using a semantic web and ontology approach. Tags structuring approaches can be considered as lightweight or heavyweight. A lightweight approach utilises dynamic information obtained from the tagging activity and is less formal. A heavyweight approach requires formal annotation scheme and data backbone to achieve the task. Section 2.6 discussed the related work on word sense disambiguation which is closely related in our research to identify the meaning of a named entity term using a knowledge base.

3.8.1 Lightweight approach

Lightweight structuring methods derive information from the tagging usage. For instance, “pasta” and “lasagna” appear frequently with “Italian food” together, it will then be concluded that “pasta” and “lasagna” and “Italian food” are related to each other. The output of the structure are based on usage statistics and The characteristics of such an approach is that structure is generated dynamically based on the relationships between the tags but the relationships between tags are undefined and require interpretation (i.e. to identify “pasta” and “lasagna” as kinds of Italian food).

Tim Berners-Lee defines the Semantic web as an extension of the World Wide Web which contributes to the evolution of community knowledge as a whole [82]. By utilising key enabling technologies to promote interoperability, a web is to be considered as a universal medium for information exchange through formal definition of terms, concepts and relationships within a knowledge domain.
Her et al. considers the Semantic Web and Web 2.0 as two complementary paradigms which can be converged and merged eventually [83]. Before ontology can be automatically extracted from Web 2.0, a semi automated approach which abides by the following guidelines will be able to achieve a smooth and user-friendly transition from Web 2.0 to a semantic web (Table 3.2).

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-automated tagging</td>
<td>Utilises text classification techniques and provides assistance to users to guide them towards consistent tagging, hence overcoming common tagging problems</td>
</tr>
<tr>
<td>Tag merging</td>
<td>Provides techniques to group synonyms and misspelled tags together for a more consistent tagging and increases the annotation quality</td>
</tr>
<tr>
<td>Related tags identification</td>
<td>Utilises statistical information such as tag co-occurrence to identify relationships between tags</td>
</tr>
<tr>
<td>Tag rating</td>
<td>Allows users to rate tags and distinguish high quality tags for use by domain experts in updating the core ontology</td>
</tr>
<tr>
<td>Information extraction</td>
<td>Employs existing knowledge base to identify the relationship between entities and connects with the ontology</td>
</tr>
</tbody>
</table>

Table 3.2 Guidelines for combining Web 2.0 and Semantic Web

Angeletou et al. proposed a Folksonomy Ontology Enrichment (FLOR) model to enrich Flickr tags using the automatic WordNet [11]. FLOR utilises a three phase approach, namely; Lexical Processing, Sense Definition and Semantic Expansion and Semantic Enrichment to achieve semantic annotation. The function of Phase 1 is to identify and exclude meaningless tags from the tags collection (e.g. special characters, non-English and number tags). The filtered tags will then be normalised to expand the Lexical Representation coverage (e.g to expand santabarbara into santa_barbara, santa+Barbara, santa-barbara, etc). The second part of the algorithm applies WordNet and tag co-occurrence information to determine the tags meaning and further increase the tag set. For example, by identifying apple as co-occurring highly with mac, apple can be recognised as the company and the tagset {apple, mac} can be further expanded into {apple, mac, computer, macOS}. Lastly, the final tagset is returned by querying the semantic web.

Similarly, Laniado et al. employed WordNet to add explicit semantics with a static hierarchy of concept [10]. A semantic ontology tree of related tags is built from selecting N bookmark with M amount of frequent tags used for the site. A hypernym chain is then created as a path to the root of the term hierarchy and then merges with the existing tree. The tree is then
rearranged by removing popular terms in the “black list” defined and pruned by removing any tag with a branch less than $K$ (default value is 2). The tree is then re-ordered by its weight and integrated into the existing delicious website as a navigation interface to guide the users in browsing the related tags hierarchically. Prototype implementation has been demonstrated and shows that by browsing the tag “pasta,” users can traverse along its sibling nodes such as *macaroni*, *gnocchi* or other related terms such as *chicken*, *seafood* and *vegetable*.

### 3.8.2 Heavyweight approach

Ontology is a formal method to define and model the relationships between concepts and objects within a domain of discourse, which is widely used in the computer science area to facilitate the formal information exchange. Active research suggests the use of ontology as the main backbone to govern the hierarchy of resources and manage the tags. Ontology construction requires professional input to build the hierarchy, maintain the dictionary and perform manual indexing. Users are required to receive prior training in order to understand the whole ontology before being involved in classification.

A heavyweight structuring approach aims to present a formal and well-structured taxonomy for the social tags. For instance, information experts can utilise ontology to model real world objects using formally defined types, properties and relationships. Tags can be utilised as part of the concept of ontology or to enrich the metadata from external sources. Every tag will need to be attached to an ontology entry to explicitly represent its usage. For instance, in food ontology [84] “pasta” is defined as “ConsumableThing $\rightarrow$ EdibleThing $\rightarrow$ Pasta” and it can be extended and used as the superclass of “FettucineAlfredo”. This information can be written in a formal language such as RDF and is used to perform data exchanges with other data sources.

Gruber defines ontology as an explicit specification of a conceptualisation to describe the commitments of a set of agents in a specific community so that they can communicate without operating on globally agreed upon terms. In the article “ontology of folksonomy: a mash-up of apple and oranges” [85], Gruber demonstrates that folksonomy and ontology are not opposed to one another and that both techniques should converge and complement each other. The folksonomy is hence defined as $Tagging = \{object, tag, tagger\}$. This set represents the relationship between the object, the tags and the user who annotates the object and is useful within a singular system, however, in order to facilitate ontology sharing and to indicate the quality of tagging (for spam elimination or error detection), the tagging model should ideally be $Tagging = \{object, tag, tagger, source, polarity\}$ to indicate the source of the tagging tuple and accuracy using polarity (i.e. + or -).
Passant [12] and Echarte et al. [17] both proposed to use ontology as a fundamental to maintain and manage folksonomy. Passant’s motivation emerges from the needs of indexing blog posts in EDF R&D (a blogging platform that allows engineers and researchers to exchange information). Users are encouraged to provide meta-data either from predefined categories or free keywords. He proposed to apply domain ontology on the existing folksonomy and uses SIOC (Semantically Interlinked Online Communities) which are based on RDFS (RDF Schema). Users are required to validate the blog post and associated tags whenever a new blog entry is posted. They will then be required to associate the tag with their relevant concepts. The research project also incorporates tag clouds as the main tools for visualising the tags and providing suggestions to user. The preliminary experiment shows that this approach efficiently gets rid of various tagging classification issues.

Echarte et al. proposes a similar approach that uses ontology to formalise and define the domain vocabulary [86]. A set of ontological properties and algorithms has been proposed along the way of the model formalisation to specify the method to transform existing folksonomy into a described model. Delicious is selected as the testing website and the experimental results showed that the proposed model efficiently responds to tags variations with the following properties (Table 3.3).

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Source of the application</td>
</tr>
<tr>
<td>Resource</td>
<td>Type of content (i.e. documents, photos, bookmarks)</td>
</tr>
<tr>
<td>User</td>
<td>Individual tag creator or owner who performed the tagging</td>
</tr>
<tr>
<td>Tag</td>
<td>A collective property which represents the tag concept, variations (synonyms and spelling) and nature of the tag (personal or common)</td>
</tr>
<tr>
<td>Annotation</td>
<td>The relationship between user action, tags and the resource</td>
</tr>
<tr>
<td>AnnotationTag</td>
<td>The relationship between annotation and tags</td>
</tr>
<tr>
<td>Polarity</td>
<td>The quality of the tag (positive or negative)</td>
</tr>
</tbody>
</table>

Table 3.3. Ontology properties for folksonomy formalisation

The study demonstrates its ability to collect similar tags into a group (e.g. \{Ruby, Ror, Rails, RubyOnRails\}) which significantly reduced the amount of tag class required for the dataset. The reduction in number of tags by grouping different tags variations improves the overall quality and prevents problems being detected in tag cloud visualisation. The research also aims to make use of the new modelling technique to represent any kind of folksonomy and offer
Mika formally defines folksonomy at an abstract level using a ternary edges annotation \( T \) as a subset of \( A \) union \( C \) union \( I \) where \( A \) represent actors (users), \( C \) represent concepts (tags) and \( I \) represents an instance (objects) [87]. The study discusses the changes in community commitment and evolution and the entering and leaving of members causing changes to the ontology. Therefore, an easily scalable and maintainable Semantic Web is required to adapt to these changes. Co-occurrence information is captured to build related graphs and network analysis is applied on the tags later to discover semantically related tag clusters. Two emerged ontologies: concept relations based on instance overlap and concept relations based on actor overlap can be obtained by folding ternary relations into binary. This research represents one of the important works in deriving emerged semantics from collating results of individual actions in a community. Hamasaki et al further extends Mika’s approach by including an actor-actor dimension in their Homophilic Actor Mediated Activation (HAMA) model to show that introducing relationships between actors is able to improve the performance by providing wider coverage [88]. The P-Concept (personal concept) is introduced to detect the concepts conceived in the community before the sharing process to provide a cleaner overview for web conceptualisation.

While the heavyweight approach can effectively model the data formally, we consider that such an approach is not appropriate for use in this project. A heavyweight approach requires users to understand the ontology model prior to performing tagging, and users are required to provide additional information, which eventually changes the tagging process. In a dynamic environment such as online news, a lightweight technique is considered the preferred approach as it is dynamic and requires less computational effort. While the output information is not as rich as in a heavyweight algorithm, it is sufficient to distinctively differentiate between the individual meanings of the tags, and the data can be enriched at a later time using other systems.

### 3.9 USER STUDY ON TAGGING

The concept of tagging and using keyword to label resources has been widely used since the past decades. For example, web page creators seldom provided META tags in their webpages until they were aware of the effect of ranking their webpages in search engines. Ames and Naaman [89] conducted a study on Flickr to interview 13 participants who take photos with a digital camera and mobile phone and upload to Flickr. The study results showed that users primarily tag the photos for retrieving purposes and social sharing purposes and those users are willing to annotate content when more incentives are given and they are found to benefit from the
tags suggested. In certain cases the suggested tags are not directly applicable; however it still provides an indicative guide for users to indicate what nature of tags is appropriate.

Bar-Ilan et al. conducted a case study [90] to provide an insightful study of structured and unstructured tagging. A group of 47 students were selected to assign tags to 12 images related to Israeli and Jewish cultural heritage. For structured tagging, users were asked to provide tags according to a set of structured forms, including the following fields: general themes, symbols, personalities, description of event, location of event, time of event, object type, object creation date, creator, related links and additional information. Students who participated in unstructured tagging had the freedom to give a list of free-text keywords to describe the content of the pictures and were encouraged to put in keywords related to artistic aspect. The results showed that structured tagging produces a more detailed description, however there is no such generic structure that is applicable to all the cases, and the forms started to become complex when attempting to capture more data. Furthermore, complex form results discouraged users from providing descriptions. The other problem that arose was that some values can be assigned to several fields and each tagger has their own decision to allocate the value to a different field. This causes confusion and complicates the discovery and retrieval process. The author hence suggested that using free tagging that does not limit the images, appears to be a better approach to apply.

Carmagnola et al. presents a study on a tag-based user model to investigate how users benefit from tagging [91]. Thirty-nine users were selected to participate in the study, which showed that 76% of those users tagged events using proposed tags (tags from the event description) and 24% of the users tag events from a freely chosen keyword. These findings further support and confirm the idea that most users will accept and be benefited by the tag suggestion provided by the system.

3.10 TAGS VISUALISATION

A main requirement of a social system is the functionality to visualise the tags. Tags visualisation allows users to capture the vocabulary of the social community easily through a visual depiction. Tags can be visualised in the following presentations:

- **Tag index** – The tags are presented in their alphabetical index. No importance or frequency of occurrence is illustrated. No pagination is provided in this visualisation method but users are allowed to explore the entire tags collection.
Flat tag list – This is the basic visual depiction of a tag collection where important tags are displayed and sorted alphabetically. The importance is determined by the frequency of occurrence and the number of tags is decided by the system however the weights of tags in the tag cloud are not exhibited. The size of the list are relatively smaller than a list (within 20 – 30 tags) which allow users to easily obtain a quick snapshot of the current hot topics.

Tag cloud – The widely adopted visualisation approach in most of the social bookmarking and tagging enabled site. Frequently occurring tags are represented in a different font size and font weight therefore users can easily identify the topics that are currently of greatest interest to other users.
As part of the study of folksonomy, researchers realised that when the amount of tags increased, users started to face difficulties in visualising the tags and the need for tag visualisation schemes emerged. One of the main requirements is an effective scheme to visualise the tags efficiently. The main drive behind this need is to investigate the effect of the different methods that can be utilised in presenting the tags.

A tag cloud is a popular approach used for visually illustrating a tag list. It is a collection of tuples consisting of resources and links [92], which aims to enable the user to visualise important topics and tags at a glance, with some kind of visual aid to represent the tag’s popularity. This is usually achieved by larger font or another emphasis such as contrasting colour and variation of font weight.

The first appearance of a tag cloud on a popular website was on Flickr. The implementation of it is also seen at other popular collaborative tagging sites such as Delicious and Technorati. Tag cloud implementation receives a good evaluation and holds wide acceptance by most of the tag based systems. A Hassan-Montero and Herrero-Solana point out that the current tag cloud model does possess a few weaknesses [93]. The frequency of appearance might not be a suitable solution to generate a tag cloud and frequently used terms are not always good discriminators. Alphabetical arrangement of tags does not infer any level of semantic relationship between tags. They also argue that even though folksonomy is commonly defined as a flat space of keywords, the visual presentation can be further improved if associative and semantic similarity between tags can be modelled. The proposed algorithm adopts the K-means clustering method to improve the traditional tag cloud and the algorithm improves the tag cloud by reducing the number of less significant tags and at the same time groups semantically similar tags together.

Inspired by Hassan’s research, Kaser and Lemire [94] agreed with Hassan-Monterero that the associated tags should be placed close to each other and identify two problems of tag clouds, which are space wastage and large clumps of white space. The papers present an approach using the Electronic Design Automation (EDA), min-cut and floor planning to efficiently optimise the layout of tag clouds. The results of the experiment show that the algorithm works well, however there are some limitations caused by the design of HTML and CSS and expectantly future development of CSS3 will be able to help to address these issues efficiently.

Dubinko et al. from Yahoo research, proposed a novel method to help the users in visualising the tags over time [95]. The proposed approach is closely related to our current research, as temporal evolution is one of the main factors which affect the news and the tags value in the news domain. The research proposed an indicator named “Interestingness” to capture interesting tags within a given interval. A tag is said to be interesting within an interval if
it appears more frequently in the interval and also appears less frequently outside the interval, and an infrequent tag that occurs only during the interval does not imply that it is the most interesting tag within the interval. The paper uses river metaphor and waterfall metaphor to show the visualisation of tags, and the results are promising. Due to its ability to be generated easily on a broad folksonomy to represent the global tags popularity [96], tag cloud visualisation will be implemented in our system to show the effect of tags in the news headlines.

3.11 SOCIAL TAGGING FEATURE

As the popularity of collaborative tagging, as well as the number of users is increasing and has been widely accepted, the collaborative website has been able to gather valuable information about users over time.

Santos-Neto et al. [97] presented their work in to analyse whether a usage pattern can be harnessed to improve navigability. The two main collaborative scientific literature sites CiteULike and Bibsonomy were selected for the experiment. The research reports the result from a quantitative study that focuses on three main points: tagging activity distribution, user similarity and structure of the tagging community. Tag activity distribution refers to the frequency of a resource being tagged and on the tag as an important indicator to understand the entity that draws user attention. Due to their understanding of the Importance of user interest, the study presents an interest-sharing graph to capture the commonality in user interest for the whole user population. The user interest similarities can be used to model the preferences among individual users for some further possible grouping and further understanding of users. Lastly the paper presented a method to measure the navigability using the Entropy method. The study shows that an interest-sharing graph is a good basis to develop a recommendation system and that it helps to improve navigability. This research is done on scientific literature bookmarking websites and the findings show that most of the users do not share similar interests.

Wu et al. points out that a large amount of valuable information is available in the underlying tagging websites and the current design does not support the extensibility of the requirements to mine the data [98]. Their work presents a design to enhance the collaborative tagging systems to meet some key challenges: community identification, user and document recommendation and ontology generation. The design uses a spectral method to identify the global communities by considering documents, tags and users as nodes in a network and shows that links are established between the nodes. SVD (Singular Value Decomposition) is then applied to represent the prominent users and key documents related to the topics using EigenVector. To consider the user, tags and documents as they link together, the authors present a modified HITS algorithm (an algorithm known for identifying high impact sources in a hyperlinked environment) [99] to extract experts and high-quality documents related to a
keyword with respect to the usage of tags. After a quality document is found, pair-wise similarities are computed to further retrieve similar documents. Lastly, the study evaluates the results of using a hierarchical clustering method to generate the ontology for the tags.

As folksonomy has grown, Hotho et al. realised the need of a ranking mechanism for the folksonomy model, and proposed a formal folksonomy ranking scheme named “Folkrank” [100]. They define folksonomy as a quadruple dimension tuple $F$ which consists of $U,T,R,Y$ where $U$ denotes the users, $T$ denotes the tags and $R$ denotes the resource and $Y$ represents a ternary relationship between $U,T$ and $R$. A subset of 10,000 resources with tags from Delicious were selected as the basis of the dataset. Folkrank and Adapted Pagerank are selected for their performance comparison and the results show that Folkrank provides excellent results when conducting folksonomy queries for topically related elements. The authors suggest that Folkrank can be used as a recommendation generation tool to present relevant documents, related tags and users that share a common viewpoint.

In Xu et al. [101], 1 month of tagging data is used (126,304 unique tags, 690,482 URLs and 10,109 different users) from Delicious to study the methods for semantically related bookmark exploration which is not well supported in present social bookmarking services. The authors define social bookmark annotation as a set of quadruple feature consisting of the user, resource and tag. In that research, authors do not consider time as the main factor and it is omitted in the calculation. The results show that different tags that are assigned to the same resources are usually semantically related with one another and users have similar interests if they share many semantically related tags. Related resources are tagged many times by semantically related tags and users are related to each other if they semantically tagged the same resources. An Extended Bigram Separable Mixture Model is then modelled to a tripartite probabilistic model to obtain the emergent semantics in social annotation data.

With the similar approach to find out the emergent semantics in data, Aurnhammer et al. [102] proposed a method to use visual features to improve navigability and overcome various problems in tag oriented navigation. The authors identify three common problems in tagging practice: spelling mistakes, synonymy and homonymy. The research selects 3000 Flickr photos from 12 random and uses colour and texture as the visual features. The first two moments (mean and standard deviation) in RGB colour space are calculated to represent the “colourfulness” of an image and Oriented Gaussian Derivatives (OGD) are used to represent the surface and structure of an object (represented in linearity, periodicity and directivity of an image). The similarity between images is computed using Euclidean distance. Results show that the combination of tags and visual features performs better than using single features only and indicates the possibility to
use tagging on other media types such as music and video to further improve the retrieval precision.

3.12 CONCLUSION

Folksonomy is steadily gaining in popularity. We believe that folksonomy will lead the trend and having considered the arguments from Shirky [103], we are motivated and believe that folksonomy is the way to go to improve the navigability and information retrieval for use in a dynamic environment. Folksonomy is currently standing simultaneously on both the best side and the worst side of information organisation. The freeform nature of the tagging process causes folksonomy to be uncontrolled and chaotic. Various issues such as synonym, homonym and ambiguity are yet to be resolved.

One approach to structuring folksonomy is to utilise ontology. Ontology is an effective technique for managing resources and content within domains with fixed vocabulary sets and less frequent varied hierarchy; however the construction cost is expensive [86]. Though automatic ontology construction solves the problem of manual labour, scalability issues remained unsolved. While being applied to folksonomy, ontology can help to manage and foster strong relationships among users, objects and tags but most of the methods reported in the literature incurred additional user intervention which causes the user to change their usual tagging habits. However, such changes and involvement in structuring hierarchy does not generate any immediate incentive for them. In addition, implementation of such a heavyweight approach might deteriorate folksonomy system performance and eventually become one of the entry barriers that refrain public users from participating in tagging.

On the other hand, as folksonomy represents the social aspects in the web dimension, it should be noted that by nature contents in social tagging are dynamic and the value of the content is often ephemeral and evolved swiftly over time. Therefore, it is not cost effective to engage a professional just to maintain the hierarchy, construct the terminology and perform classification for such an application. Therefore the combination of a lightweight indexing scheme and free form tagging will be a good solution which balances both sides of the weight. The rigid structure of structured annotation and the free form nature of social tagging complement each other and facilitating both content classification and the information retrieval process. In the front end, the user will be able to perform tagging and browsing activities while the tags are being controlled and managed behind the scene. The seamless bridging of the indexing scheme and unstructured annotation will result in a more promising system and improve the overall quality and reliability of information retrieval. In addition, the popular tags can be identified easily without going through complex indexing processes and the user is able to visualise the tags. Controlled tags can then be used as a guideline during the tagging process.
Various literatures have also pointed out the benefits of folksonomy, such as the improvement of recall in information retrieval as the tags always reflect the vocabulary used in the community and the most popular keyword implies the content that is of wide interest to others. The effect of this approach is especially valuable in the dynamic domain for its ability to reflect the topics shared and of interest to the reader. The tags also reflect the majority understanding and viewpoint towards the content and generate results in a confined space. We envisage that by resolving ambiguity and synonymy, the folksonomy result can be further improved and enhance the overall retrieval performance and users experience. In addition, social tagging has been identified as the suitable approach for lightweight cataloguing [105].

What is lacking in the current online news system is a simple but effective mechanism to provide further interaction between the users and the system. Social tagging appears to be a viable solution for use in an online news website as a primary tool to facilitate news clipping, news organising and news sharing. The tagging functionality can be built on top of the rich amount of metadata from the news article, hence providing users with an additional feature to complement keyword based information retrieval. Additional work is still required in order to strengthen various outstanding issues that exist in folksonomy and the utilisation of the named entity nature of news articles to improve the overall system performance.

In this chapter, we identified the key work that has been conducted that relates to tag structuring and named entity indexing. Tag structuring is important to structure folksonomy, which can be used to fix various problems therein. There are lightweight and heavyweight approaches. Both approaches aim to identify the meaning of the tags. The lightweight approach connects tags with Web 2.0 structured information and the heavyweight approach defines formal properties, classes and relationships between the tags.

There are still limitations existing in the previous works. Heavyweight approaches provide formal and rich information regarding the tags, which enables the tags to be semantically rich in information. However, such an approach is laborious and requires expert judgement. This requires the users to receive prior training before being able to utilise the system effectively. In news domain, such additional processing and cognitive cost appears to be cost ineffective. On the other hand, lightweight approaches by nature are more dynamic and suitable for such application, however, current work focussing on structuring named entity tags is still limited.

In the next chapter, we present the framework of our system which provides a high level overview to aid in understanding the architectural logic of the whole system.
Chapter 4: System Framework

4.1 INTRODUCTION

As discussed in Chapter 3, named entity (NE) indexing approaches can be utilised to improve information retrieval. The NE index enables users to obtain richer information while consuming news feeds and to perform complex queries which complement traditional terms based searching. To further strengthen the application of the NE index and to overcome inherent linguistic issues such as ambiguity and synonymy, the NE index can be linked to knowledge. The following section describes an architectural perspective of our framework as depicted in Figure 4.1, the work presented in this thesis covers the main components required for constructing a named entity index and the web news system.

4.2 NEWS INDEXING

Chapter 5 describes the process to continuously monitor online news streams and download content regularly for indexing. The indexing process comprises of the following components:-

4.2.1 Indexer

The news collector is the primary component that gathers news document HTML files from news publisher sources automatically within a specific duration. The collector will need to support major feed types such as Rich Site Syndicate (RSS) 2.0 [106] and ATOM publishing protocol 1.0 [107]. Meta information such as the source, publication and publication time are collected. The HTML tags are stripped and forwarded to NE tags extraction. Creator tags and user tags are then extracted from the document source and NE tagger tools are applied to extract named entity tags from the content. The pre-processor proceeds to check and ensure the tags are well-formed and then commits the data into the database.

4.2.2 Entity extractor

The entity extractor component functions to extract named entities from the news content.

4.2.3 Named Entity Mapper

In order to bridge the gap between free form tags and structured data, the meaning of the tags will need to be explicitly identified [108]. The NE mapping component in the framework is designated to perform such tasks. The two main operations in the NE mapper are sense detection
and sense disambiguation. Sense detection functions to identify the concise meaning of a tag and sense disambiguation resolves ambiguity between terms that share the same surface form but with a different meaning. The root form of a tag and relevant surface forms are detected to ensure a singular data version presence. A knowledge base is required to provide the necessary information to support such operation. The NE mapper then decides on the most suitable sense for the tag and the root forms unique identifier can be extracted and be used to construct a controlled vocabulary dictionary.

4.3 NEWS SYSTEM

Chapter 6 describes the system that utilises the NE index and provides a platform for users to browse and retrieve information.

4.3.1 Graphical User Interface

The Graphical User Interface (GUI) of a system is the main entry point which provides a user-friendly and easily accessible interface that facilitates the interaction between users and the system. The user interface provides easy and intuitive navigation with the aid of rich visual elements. The visual design adopts a simple and straightforward approach which aims to integrate additional features without radically changing the users browsing experience. As most of the operations (e.g. retrieving entity list, retrieving tag suggestion and validating tag sense) required multiple round-trip data transfer and post back between the server and the client, asynchronous client-side scripting such as AJAX (Asynchronous JavaScript and XML) should be used whenever applicable to provide maximum responsiveness and avoid unnecessary post back.

4.3.2 Visualisation

The browsing component provides support to assist users in navigating the site. NE index functionality is maximised to provide reading time support with different presentation methods such as relevant named entities and tag clouds. The browsing component utilises statistical information from tags usage and voting to generate dynamic information that provides serendipitous information discovery that suits the nature of news reading.

4.3.3 User management

The user management component is responsible for managing users and their activity in the system. User information including user name, password and friends lists are maintained. A friend’s interface is provided to moderate the friends process and all activities are tracked so that they can be shown and shared with other users.
4.3.4 **Organisation**

This module provides tagging support to the news readers. Every registered user is allowed to freely assign tags to the articles. The tagging is not limited to creator only, but opened to the public in order to promote a collaborative effort. Every tagging instance will directly contribute to the outcome of the whole tagging space. The tagging functionality is to be an additional feature on top of the browsing system and should not alter usual browsing activity and or introduce any major change in the diffused browsing interaction pattern.

4.3.5 **Retrieval**

For facilitating usage, an interface is also provided for users to perform NE based queries in which the users can specify the NE category from which they seek to retrieve news articles or related entities that they are interested in. The retrieval component will be in charge of tracking user search criteria and results in order to allow users to reuse at a later time.

![System architecture diagram](image-url)

Figure 4.1. System architecture
4.4 EXTENDED FUNCTIONALITY

Figure 4.5 depicts the user interaction diagram which indicates the additional functionality that is enabled by the framework as compared to the current system. The NE index enables users to read news and read about the list of entities that appears in the news articles. The five main new functions that are provided are:-

Named entity retrieval: To retrieve news articles using the combination of keyword and the designated meaning.

View related entity: To view the entity that is related to the news articles.

Friends interface: To allow users to mark other users as friends and to browse their profiles, reading history, activity log and vice-versa.

Tagging: To allow users to assign tags to resources and perform tag based retrievals.

Create clipbook: To allow users to create a manual or automatic clipbook that helps to keep track of the developments of a news topic.

![User functions in reading news](image)

Figure 4.2. User functions in reading news
Chapter 5: Tags Indexing and Structuring

The Named Entity (NE) tags indexing approach has been proven useful and its application has been widely demonstrated. Specifically in the online news domain, NE indexing has been proven useful in providing users with additional control over information such as enabling users to search through a collection by selecting their own topic of interest based on the person or company that they are interested in.

Previous studies such as that by Neptuno proposed to use ontology to archive news and provide semantic search functionality to users. We argue that such a heavyweight approach is inappropriate as the cost of ontology construction and maintenance is expensive and laborious. Users are required to receive prior training before consuming such a service. Given the ephemeral nature of news articles, such high financial investment does not generate maximum benefits for the users. Effective NE indexing on a news article can help users to quickly understand the contents of an article and the nature of that content.

Current retrieval techniques provide users with an interface to retrieve a set of documents that match a given query, but the ability of answering queries is limited. The primary step to provide users with more control in the content delivery is to construct a semantic index which represents the contents of the news article. Effective coverage of named entities in the news articles will enable users to have a better overall understanding of the content inside without having to read through the whole article. Together with the temporal information, a quick summary can be drawn to understand the content. For example, we can predict an article with “John Howard” and “Kevin Rudd” during September 2007 is related to the Australia Federal Election and an article with “Kevin Rudd” and “Wayne Swan” during Feb 2008 is related to the 2008 federal budget.

Rich metadata can be extracted from the news articles and be structured to form the semantic of the news articles. This chapter focusses on processing three main types of news metadata: name entity (NE) tags, creator tags and users tags. The extracted tags are mapped into their respective sense in the knowledge base (KB) and its URL in the KB is used as the unique identifier for data interchange.
5.1 DESIGN

Figure 5.1 depicts the architectural design of the indexing process. News articles are first collected from the source through the news publisher RSS feed. Metadata information including publication date, news source, title and description are extracted and indexed into the database.

Creator tags and user tags are then extracted from the HTML code using regular expressions and NE tags are extracted using an automatic extractor. Extracted Tags are then forwarded to a pre-processing component for error correction by fixing spelling, capitalisation and separator errors. This is done by submitting a query to Google using the NE and performs corrections if a suggestion is available (examples shown in Table 5.1).

The subsequent step is to retrieve a meanings candidate list is suggested by the expert. The expert provides suitable sense in the Knowledge Base (KB) context to facilitate the subsequent operation. The system then detects the sense of the NE in the KB and performs disambiguation if multiple senses are detected, deciding the most suitable sense for the entity. The information of the NE and its uniquely identified entry in the KB will be stored in the dictionary. The dictionary will be presented using the industry standard format in order to facilitate information exchange and interoperability.

The system then computes the Term Frequency and Inverse Document Frequency (TF-IDF) of each named entity. The TF-IDF metric is a common metric in the Information Retrieval (IR) field to indicate the importance of terms with respect to the overall corpus. Term frequency (TF) indicates the frequency of the term \( t \) occurred in the document \( d \). Inverse document frequency (IDF) represents the logarithm of the ratio between the total documents in the collection \( dN \) and the count of documents where the term appears \( dF \). The logarithm is used to normalise the effect of the document length to represent the global importance of the term. The term weight score is calculated using the following equation: - (Are you using Microsoft equation?) if it’s not your equation, reference is needed.

\[
TF = t f_{t,d} \tag{1}
\]

\[
IDF_t = 1 + \log \frac{dN}{dF} \tag{2}
\]

\[
Score = TF \times IDF \tag{3}
\]
5.2 NAMED ENTITY EXTRACTION

Named entity extraction (or Named Entity Recognition, NER) is an important task in an Information Extraction (IE) task to perform document and content analysis. It identifies the position of an atomic element in the text and classifies entities into categories [110]. An effective named entity extraction will function to understand the content of the article etc. Therefore, a high performance entity extractor is required in this project to perform such a function.

Calais [28] is selected as the named entity extractor in this research. Calais is an NE tagging tool provided by Thomson Reuters to support content interoperability and facilitate semantically rich application development. Calais utilises state-of-the-art Natural Language Processing (NLP) and machine learning techniques that automatically annotate the input document with the detected entity. Table 5.2 illustrates the classification model of utilising...
Calais for our system. Calais is capable of processing documents from raw text, HTML or XML format and produces an output of the entity detected in the source with their occurrence detection and the relationships among the entities. This nature makes Calais the most suitable candidate for use in this project for the entity detection task. The Calais classification is effective in understanding the entity nature but is not appropriate for use in the category as it is too fine grained. Based on the nature of the news article, Calais entity typing is further grouped into five categories, namely; person, organisation, location, event and others to match the nature of news articles and prevent overly fine-grained entity types being presented which might potentially cause confusion for the users.

![Unannotated Input (News / Wikipedia article)](Calais)

Figure 5.2. Calais named entity extraction and classification model

## 5.3 KNOWLEDGE BASE

Wikipedia is selected as the KB in this project as it is regularly reviewed by the public for the nature of the given information. KB candidates such as WordNet and KIMO provide well maintained structural hierarchy, but the evolvement and update frequency are relatively lower as compared to Wikipedia. In a fast changing domain such as online news, the meaning of a named entity changes instantaneously whenever it is applicable. As reported by Bhole et al. [111], meaning of a Wikipedia URL does not vary over time, which means it is safe to map an NE tag to a Wikipedia URL without regularly validating the wiki page content if the meaning of the NE changes. For example, wiki page “NBA Playoff” consists of information about the format and the history of an NBA playoff and when the information grows and evolves, a new wiki page will be created to represent its recentness (for e.g. “NBA Playoff 2008”) instead of continuously editing the original page. This maintains the consistency of the Wikipedia URL without losing its ability in maintaining currency.
5.4 EXPERT SUGGESTION

Previous studies from Cucerzan [113] and Bunescu [114] utilise Wikipedia disambiguation page structure to identify the sense of a term by selecting the page with the highest similarity with the source document. These methods present two major limitations: Firstly, the disambiguation algorithm does not eliminate irrelevant senses by considering additional information (such as term popularity and recentness). This can potentially incur additional processing time by computing unnecessary information. For example, the term “Apple” is referring to Apple Computer during the iPhone launching period but not referring to the fruit in technology news. Similarly, “George Bush” is likely to be referring to the recent United States President George Walker Bush and is less likely to be his father George Herbert Walker Bush or the NASCAR driver George Bush during the 2008 United States Presidential Election period. Secondly, even though it is not reported, we predict that these models are unable to perform accurately for other NE tags (such as user tags) as it is designed to process NE’s that appear in the document itself. For instance, the term “Democratic Party” is used to annotate an article about Barack Obama (then Presidential nominee for the Democratic Party) but the term itself does not exist in the article.

An expert concept suggestion component is therefore implemented before the disambiguation process to address the above mentioned limitations. The component reduces incurring processing time by filtering off irrelevant senses and also providing mapping support without limitations to NE’s in the document only. Search engines are selected as the experts for their capability in retrieving relevant and interesting results from a user perspective. For example, Google ranks the importance of a webpage by recursively weighting the website using global hyperlink structure [115] and Yahoo ranks search results based on the statistical information from the web directory and the usage of keywords [116]. This shows that the result from a search engine is constantly reflecting the term popularity at the time the query is conducted.

Three search engines namely: Google, Yahoo and MSN Live, were selected as the experts in this project. A preliminary experiment reveals their ability in addressing the above mentioned drawbacks. Top 20 search results from searching for the term Apple from search engines were studied, and the result shows that an average of 97% of the search results were related to Apple Computer. This means that from the search engine perspective users are more interested in Apple manufactured technology products instead of the fruit at the time the search was performed. The second experiment was conducted to verify the search engines ability in retrieving wiki page from Wikipedia. A collection of 100 tags is randomly selected from the dataset and 98.5% of the tags can be successfully retrieved via the corresponding Wikipedia page in the top 20 search results. This shows that the selected search engines are capable of returning
relevant Wikipedia pages and this information can be utilised to facilitate WSD process by reducing the scope of comparison.

<table>
<thead>
<tr>
<th>Tag Name</th>
<th>Tag Type</th>
<th>Google</th>
<th>Yahoo</th>
<th>Live</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>NT</td>
<td>Barack_Obama</td>
<td>Barack_Obama</td>
<td>Barack_Obama</td>
</tr>
<tr>
<td>Gold Coast</td>
<td>NT</td>
<td>The_Gold_Coast</td>
<td>Gold_Coast, Queensland</td>
<td>Gold_Coast, Queensland</td>
</tr>
<tr>
<td>USA</td>
<td>NT</td>
<td>United_States</td>
<td>United_States</td>
<td>United_States</td>
</tr>
<tr>
<td>rudd-kevin</td>
<td>CT</td>
<td>Kevin_Rudd</td>
<td>Kevin_Rudd</td>
<td>Kevin_Rudd</td>
</tr>
<tr>
<td>Brisbane-4000</td>
<td>CT</td>
<td>Brisbane_Central_Business_District</td>
<td>Brisbane_Central_Business_District</td>
<td>Brisbane_Central_Business_District</td>
</tr>
<tr>
<td>Giants</td>
<td>UT</td>
<td>Giant(mythology)</td>
<td>New York Giants</td>
<td>New York Giants</td>
</tr>
<tr>
<td>Guantanamo</td>
<td>UT</td>
<td>Guantánamo Bay</td>
<td>Guantánamo Bay</td>
<td>Guantánamo Bay</td>
</tr>
</tbody>
</table>

Table 5.2. Named entity tags and their expert suggested meaning Wikipedia title (key: NT – Named Entity Tag, CT – Creator Tag, UT – User tags)

5.5 SENSE MAPPING

The inherent deficiency in the current systems exists in the natural evolution of language and various issues such as polysemy, synonym and ambiguity are not yet fully resolved. One of the main challenges is to understand the underlying semantics of entity within the contextual domain. The study of entity mapping and disambiguation in specific application domains such as geography [117], Japanese language [118] and history [119] has achieved great success and indicates that efficient NE mapping can be applied to improve information retrieval performance.

Studies related to automatic large scale NE disambiguation have been conducted and the problems in identifying various surface forms detected in documents based on a reference list of words, which can be considered as a traditional WSD lexical task. Even though the reported techniques are able to extend previous work by providing wider coverage of different named entity types, we identify that there is a gap and the need exists to solve the following issues:-

a. Existing methods might fail to map and disambiguate an entity that does not appear in the source context.

b. An entity without a wiki page cannot be mapped and disambiguated effectively.

c. Unnecessary computational costs need to be eliminated.
The primary task in achieving NE sense mapping is to perform word sense disambiguation (WSD) for identifying the meaning of the tag in the given context, using an identifier in the KB. In general noun research, [120] and [121] consider word sense disambiguation as a supervised learning task where each word-sense pair is tagged manually for training a model to tag the sense of a word. In social tagging, [122] and [59] apply WordNet as the ontology backbone on del.icio.us to improve the search result, [123] integrates semantic web into folksonomy and generates a concept cluster using related terms. These systems produced highly accurate results, but their applicability and performance relied on the update frequency of KB. The training model is required to be reviewed regularly to ensure the semantics are accurate and up-to-date. Moreover, the support of NE terms provided by these methods is rather limited. Technological terms “Javascript”, “Ajax” and abbreviations “nyc”, “sgp” are unable to be recognised. To allow wider topic coverage than general nouns, [124] and [113] utilise Wikipedia to perform automatic NE sense disambiguation and [125] applied KIM ontology for NE disambiguation in news articles. These approaches provide better performance with wider topic coverage and do not require a training model, however the main drawbacks in these systems are that the system relies on the context relationships around the entity and ranks them according to the confidence score. For instance, the term “Gold Coast” can be referring to places in both the United States and Australia, if the term co-occurs with Australia, it will be considered to refer to the famous tourist attraction Gold Coast in Queensland, Australia. The performance is expected to be affected if such co-occurrence information is absent. We argue that this method might not be appropriate when performing tag mapping, as the tags do not necessarily appear in the source content. For example, the tag ‘presidential-election-2008’ is used to annotate the news “Hillary Clinton aims for the fences” [126] to indicate that the news is related to the United States presidential election event in the year 2008. However, the system might fail to detect and map the tag correctly to the sense in the KB entry, as the term ‘Presidential Election 2008’ does not appear in the article. In addition, to solely rely on the Wikipedia facility is not adequately appropriate in certain cases. For example the term ‘Olympics’ when normally used can refer to the general description of the Olympic events (such as origin, history and movements), however, it is intuitively better to map the tag as referring to ‘2008 Summer Olympics’ in Beijing during August 2008 instead of other general definitions.

We utilised expert suggestions to accelerate the WSD process and to address the above limitations. The list of suggestions provided by the experts represents the candidate senses of the NE at the time of retrieval and hence reduces the WSD processing time by avoiding irrelevant sense computation. The method extends previous approaches by including expert information to enable mapping support to more NE tag types (i.e. not only restricted to NE presence in the document).
Previous research has reported a wide range of techniques that can be used to understand the underlying semantics of a tag in social tagging. For instance, [127], [128] and [129] attempt to understand the meaning of the tags by deriving the interpretation of tags using network analysis techniques. Laniado et al. [59] maps the social tags on WordNet ontology by disambiguating the tags using a SenseRelate metric and reports the need of ontological expansion to support newer terms and proper nouns. To provide better support and effectively disambiguate named entities, researchers turn to using Wikipedia as the KB for its wide coverage on NE’s. The most similar works to date are Bunescu [114] which focus on person name disambiguation using 55 word entity centred context methods and Cucerzan [113] who uses relevant contexts based on the first paragraph of the target page and the corresponding page, referring back to the targeted entity to perform a WSD. Both approaches remove entities from the document to eliminate irrelevant context and utilise relevant information around the entity only. We argue that named entities appearing in the candidate senses page are useful in identifying the possible senses, and that comparing the co-occurrence of all NE’s is necessary in measuring similarity.

Figure 5.3 demonstrates a paradigm to support this argument. The figure illustrates the process to disambiguate the meaning of the surface form “Oscar” between “Academic Award” (a yearly entertainment award event) or “Astronotus Ocellatus” (a type of fish) in the news article “Obama asks Hudson to sing anthem” [130]. The “Academy Award” page contains three NE’s that appear in the original document while in “Astronotus Ocellatus” page, there is no co-occurring entity. Therefore, it is more likely that the term “Oscar” used in the original article is referring to the meaning “Academy Award”.

Figure 5.3. Paradigm of tag sense mapping

Figure 5.4 illustrates the logical process of the sense detection and sense disambiguation process which takes in a list of possible senses provided by the expert and outputs a candidate Wikipedia page title of the given NE. If the experts failed to provide any suggestion, the system will attempt to retrieve a valid wiki page from Wikipedia and extract every candidate sense page from the disambiguation page if the NE denotes multiple meanings. If not, the title of the wiki page or its designated redirect page will be selected as the unique identifier. The system
disambiguates the senses by measuring the similarity between the source document and each candidate sense and selects the page with the highest similarity as the designated sense. We adopt a bag of word approach which regards every article as a bag consists with NE in the document. The bag of words in a source document will then be compared with each candidate wiki page and the similarity between the source article and the candidate wiki page is measured by calculating the similarity. We selected the Jaccard index as the similarity measurement metric. The Jaccard index is the similarity metric which can be used to represent the similarity between two objects using the following equation:

\[
Jaccard(D,W) = \frac{D \cap W}{D \cup W}
\]  

(4)

where, \(D\) denotes the named entities in the original document and \(W\) denotes the entities in the wiki page. The similarity indicates the ratio of the entities appearing in both pages against the total number of entities in both articles. The wiki page with the highest similarity score will be considered as the designated sense of the NE.

![Figure 5.4. Sense disambiguation process flowchart](image)
Algorithm I - Sense disambiguation algorithm

```
SenseDisambiguation()
SET tempScore = 0
FOR every Wikipedia page in disambiguation page
  Extract entity for wikipedia page
  SET score = Similarity score between Wikipedia page and original article
  IF (score > tempScore)
    SET tempScore = score
    SET rootform = Wikipedia page title
NEXT
```

5.6 DICTIONARY

The tag mapping result is then utilised to construct a flat non-hierarchical dictionary using NE’s extracted from news articles. The aim of the dictionary is to systematically present the indexing result and also to be available for use as a controlled vocabulary, or linked to external data sources. As the dictionary focuses on NE’s, generic terms in the English dictionary such as adjectives, adverbs and common nouns will not be covered in the scope. The dictionary is designed according to ANSI/NISO Z39.19 controlled vocabulary design guidelines [131] and provides the following functionality:-

a. Explicitly defines the meaning of a tagging instance using Wikipedia (polysemy control),

b. Provides a list of surface forms detected (synonym control),

c. Provides related terms based on using temporal information and co-occurrence statistics.

Below is the detailed explanation with examples of how the dictionary achieved the requirements defined in the guidelines.

5.6.1 Polysemy

Concisely defining the meaning of a polysemous term is the main function of a controlled vocabulary for eliminating ambiguity and improving precision. Disambiguation can effectively discriminate alternate uses of words and effectively improve retrieval precision. As shown in examples from Table 5.3, this can be achieved by describing the surface form with the full qualified name of the sense, i.e. ambiguous person name is recorded in the full name format with middle name included or generation (such as Senior and Junior), abbreviated. Organisation names are written in the fully expanded form, ambiguous location names are recorded with the
country or state where it is located, and general terms will be recorded with the term nature in parenthesis or under alternate name.

<table>
<thead>
<tr>
<th>Surface Form</th>
<th>Root Form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Amazon.com</td>
<td>An American based e-commerce company</td>
</tr>
<tr>
<td></td>
<td>Amazon River</td>
<td>The largest river in the world by volume</td>
</tr>
<tr>
<td>George Bush</td>
<td>George W. Bush</td>
<td>Current United States president</td>
</tr>
<tr>
<td></td>
<td>George H. W. Bush</td>
<td>41st United States Presidents (Father of George W. Bush)</td>
</tr>
<tr>
<td>Newcastle</td>
<td>Newcastle, New South Wales</td>
<td>The second most populated city in New South Wales, Australia</td>
</tr>
<tr>
<td></td>
<td>Newcastle-under-Lyme</td>
<td>A government district in England</td>
</tr>
<tr>
<td></td>
<td>Newcastle, Ontario</td>
<td>A town in Canada</td>
</tr>
<tr>
<td>NT</td>
<td>Northern Territory</td>
<td>A federal territory of Australia</td>
</tr>
<tr>
<td></td>
<td>New Territories</td>
<td>A extension region in Hong Kong</td>
</tr>
<tr>
<td></td>
<td>Windows NT</td>
<td>An operating system from Microsoft</td>
</tr>
<tr>
<td>Oscar</td>
<td>Oscar (fish)</td>
<td>Common name of Astronotus ocellatus</td>
</tr>
<tr>
<td></td>
<td>Academy Award</td>
<td>Annual merit award in the film industry</td>
</tr>
</tbody>
</table>

Table 5.3. Example of polysemous terms

5.6.2 Synonymy

Table 5.4 displays the example of synonyms from the dictionary. Synonym control provides functionality to identify root forms of a term and group semantically similar surface forms together to prevent redundancy and increase recall rate, by automatically retrieving related items using surface form variations. Each surface form of the term should be explicitly redirected towards its root form only to ensure singular version of content is available and provide concise coverage.

<table>
<thead>
<tr>
<th>Term</th>
<th>Surface form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>Beijing, Peking</td>
</tr>
<tr>
<td>George W. Bush</td>
<td>Bush, George Bush, George Walker Bush</td>
</tr>
<tr>
<td>National Basketball Association</td>
<td>NBA</td>
</tr>
<tr>
<td>September 11, 2001 attacks</td>
<td>911, 9-11, pentagon attacks, sept-11-terrorist-attack</td>
</tr>
<tr>
<td>Republican Party (United States)</td>
<td>GOP, Grand Old Party, GOP House, Republicans</td>
</tr>
</tbody>
</table>

Table 5.4. Example of synonyms
5.6.3 Presentation

To allow interoperability with other systems, we selected the Resource Description Framework (RDF) [132] as our schema for the data representation (Figure 5.5) with Dublin Core (DC) [133] metadata which is designed for the cross-domain information exchange.

RDF is a formal information modelling notation defined under the World Wide Web Consortium (W3C) specifications. RDF notation enables encoding and exchange and reuse of structured metadata. The key mechanism in an RDF statement is to store and exchange information and facilitates among systems using a machine readable format. Additional ontology languages (e.g. RDFS, OWL) can be built upon RDF. The reason for enabling RDF notation in the dictionary is to allow other systems to utilise the controlled vocabulary to obtain additional information and to enrich the terms in the dictionary from other structured databases.

DC metadata is a popular standard used to support cross-domain resource description. The aims of DC are to standardise the method of describing resources online to facilitate findability. The key characteristic of DC core standard is to focus on using simple descriptions to support resource discovery. DC metadata has been widely used to describe digital materials such as video [134], Images [135] and bookmarks [136]. Table 5.5 listed 12 properties provided by the dictionary which comprises of the six qualifiers from the original DC element set and six additional properties to represent a term entry.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:qut="http://www.qut.edu.au/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:term="http://purl.org/dc/terms/"
  xmlns:geo="http://www.geonames.org/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0"
  xmlns:doap="http://usefulinc.com/ns/doap.rdf#"
  xmlns:ri="http://purl.org/sa/sio/ri/"
  xmlns:sio="http://purl.org/sa/sio/"
  xmlns:sioc="http://rdfs.org/sioc/ns#"
  xmlns:siocxt="http://rdfs.org/sioc/types#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:dbpedia-namespace="http://dbpedia.org/ontology/">
  <rdf:description
      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
        <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
          <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
            <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
              <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                  <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                    <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                        <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                          <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                            <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                              <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                  <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                    <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                        <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                          <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                            <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                              <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                  <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                    <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                        <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                          <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                            <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                              <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                  <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                    <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                        <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                          <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                            <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                              <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                                <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                                  <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                                    <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                                      <dc:identifier rdf:resource="http://purl.org/dc/elements/1.1/identifier">
                                                                                      </dc:identifier>
                                                                               </dc:identifier>
                                                                              </dc:identifier>
                                                                            </dc:identifier>
                                                                          </dc:identifier>
                                                                        </dc:identifier>
                                                                      </dc:identifier>
                                                                    </dc:identifier>
                                                                  </dc:identifier>
                                                                </dc:identifier>
                                                              </dc:identifier>
                                                            </dc:identifier>
                                                          </dc:identifier>
                                                        </dc:identifier>
                                                      </dc:identifier>
                                                    </dc:identifier>
                                                  </dc:identifier>
                                                </dc:identifier>
                                              </dc:identifier>
                                            </dc:identifier>
                                          </dc:identifier>
                                        </dc:identifier>
                                      </dc:identifier>
                                    </dc:identifier>
                                  </dc:identifier>
                                </dc:identifier>
                              </dc:identifier>
                            </dc:identifier>
                          </dc:identifier>
                        </dc:identifier>
                      </dc:identifier>
                    </dc:identifier>
                  </dc:identifier>
                </dc:identifier>
              </dc:identifier>
            </dc:identifier>
          </dc:identifier>
        </dc:identifier>
      </dc:identifier>
    </dc:identifier>
  </dc:identifier>
</rdf:description>
</rdf:RDF>
```

Figure 5.5. Sample RDF entry snapshot
### Table 5.5. Field used in the dictionary

<table>
<thead>
<tr>
<th>Field</th>
<th>DC Terms</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
<td>Name of the resource</td>
<td>Academy Award</td>
</tr>
<tr>
<td>Identifier</td>
<td>Identifier</td>
<td>The unique identifier for the entity instance</td>
<td>2F1E4FC0-81FD-11DA-9156-00036A0F876A</td>
</tr>
<tr>
<td>Language</td>
<td>Language</td>
<td>Language of the resource in ISO639 format</td>
<td>EN</td>
</tr>
<tr>
<td>Type</td>
<td>Type</td>
<td>The category of the resource</td>
<td>Event</td>
</tr>
<tr>
<td>Abstract</td>
<td>Abstract</td>
<td>Short description of the resource</td>
<td>The Academy Awards, popularly known as the Oscars,........</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>The date of the RDF is generated</td>
<td>25/09/2008</td>
</tr>
<tr>
<td>News</td>
<td>n/a</td>
<td>Ten recent news the term appear</td>
<td>3F2504E0-4F89-11D3-9A0C-0305E82C3301</td>
</tr>
<tr>
<td>Image</td>
<td>n/a</td>
<td>Image URL of the resource</td>
<td><a href="http://upload.wikimedia.org/wikipedia/commons/d/dc/Academy_Award_Oscar.jpg">http://upload.wikimedia.org/wikipedia/commons/d/dc/Academy_Award_Oscar.jpg</a></td>
</tr>
<tr>
<td>StartDate</td>
<td>n/a</td>
<td>The start date for retrieving related tags</td>
<td>16/09/2008</td>
</tr>
<tr>
<td>EndDate</td>
<td>n/a</td>
<td>The end date for retrieving related tags</td>
<td>23/09/2008</td>
</tr>
<tr>
<td>RelatedEntity</td>
<td>n/a</td>
<td>A collection of related NE between StartDate and EndDate</td>
<td>Brad Pitt, Heath Ledger, Steven Spielberg, Los Angeles, United States, Hollywood, New York, Love Story</td>
</tr>
<tr>
<td>SurfaceForms</td>
<td>n/a</td>
<td>The other form variations of the resource</td>
<td>Oscar</td>
</tr>
</tbody>
</table>

Apart from simply mapping the NE tag to Wikipedia and extracting basic descriptive information, the dictionary is also able to provide a list of temporal ranked related tags based on the date scope defined by the user. This feature is practically useful in a fast changing domain such as online news as temporal information plays an important role when information evolves swiftly and the ability of ranking related NE tags will provide great value to the user by enabling them to obtain concise, relevant information. For example, during the Australian Federal Election 2007 period, Australian Prime Minister Kevin Rudd (then Labour party leader) is strongly related to John Howard (then Prime Minister) and during the 2008 Australian Federal Budget
announcement period, Kevin Rudd is strongly related to Wayne swan (Australian federal treasurer). This temporal information is beneficial to the dictionary consumer to enable them a better understanding of the relationship between named entities at different points of time.

5.7 CONCLUSION

We proposed a new approach to effectively perform tag mapping by extending the existing WSD techniques, combining the search engine facility and effectively mapping the NE tags to the Wikipedia knowledge base. The uniqueness of this model is that it does not require any prior training, which enables the system to become more flexible and be able to reflect the meaning of the NE as soon as any new senses are detected without having to reconstruct the training model. In addition, we also utilised expert suggestions to enhance the process by exploiting the mature and well-established search engine technology to assist in retrieving the meaning which is closest to the user context. By doing this, we effectively avoid spending additional computational cost in repetitively measuring the content against wiki pages which have a low popularity or are totally irrelevant at a specific point of time.

In next chapter, we present OnestopNews, a collaborative news system, which utilises the dictionary generated in this chapter to demonstrate how users can benefit from the improved and enhanced search results, through seamlessly integrating a reliable named entity controlled vocabulary as a dictionary in the data source.
Chapter 6: OneStopNews

Numerous tools such as Google News [31], Google Reader [62], Delicious [7] and Digg [32] have been actively developed in the past few years to provide support for users to obtain information efficiently. Technology such as Real Simple Syndication (RSS) feed has been enabled to automatically gather information to ensure users will not miss out on any headlines. News aggregators automatically retrieve and cluster relevant news together which provides users with a broad coverage across various news topics. Users can also post the URL of the interesting news article to their preferred social system to create bookmarks and share information with friends.

The news system today has achieved a gigantic leap from the past. While the current news websites effectively address various delivery and browsing limitations, few problems related to browsing, organising and sharing still remained unsolved. Users are still unable to read and search for news articles based on the entities, and organising and sharing facilities are not yet fully integrated into the news system. With that in mind, the proposed system is designed based on the following considerations:-

- **Retrieval** – To allow users to retrieve more relevant and concise information on top of keyword based searches.

- **Centralised social platform** – To facilitate onsite social activity as part of the system without relying on third party services.

- **Management support** – To include tools and utilities that help the reader to manage news articles more efficiently.

We present OnestopNews, a collaborative news system which aims to provide users with a better browsing and collaborative sharing platform by enhancing the current system functionality. The system integrates the news index and utilises the dictionary developed in Chapter 5 as the controlled vocabulary, which allows users to browse and organise news articles more effectively by providing NE based support. Tagging functionality is implemented using the controlled vocabulary to provide users with tagging assistance and to overcome the potential problems that might be caused by its free form nature. Numerous additional features are implemented to enhance the systems capability of organising and sharing information.
6.1 USER SCENARIOS

The users can be classified according to the following, based on their requirements of functionality and the operation required to be performed while consuming an online news system:

- **Basic users** - Basic users represent the group of generic news readers who utilise the news browsing system for learning of breaking news stories. The users in this category do not require high precision information retrieval as their main interest is to obtain a wide coverage of news stories. Therefore high recall is important.

- **Intermediate users** - Intermediate users are users who have more specific news requirements as compared to basic users. They require specific information about a certain entity (e.g. Celebrities, Politicians) or wish to follow up breaking news story developments (such as Federal Budget 2008). Simple functionality such as sharing, favourite list and voting are sufficient in supporting this operation and a good balance between precision and recall will be beneficial to assist the readers in capturing a good picture for their topic of interest.

- **Advanced users** - Advanced users utilise the news articles for specific purposes such as research. They require additional tools to support their operation such as bookmarking, tagging, sharing and news clipping to help them in managing and analysing a large amount of data.

During the users study phase, we observe the users’ behaviour and identify the following scenarios, which are commonly seen in the users while performing different task to address individual requires by the functionality provided.

**S1**: A casual news reader utilises online news as an activity to spend time and learn about current affairs across multiple genres. He does not have any specific topic of interest but would like the system to show him articles that are interesting, such as those most voted on or recommended by his friends. He would also like to discover trends about what persons or events are attracting the most public attention at the moment.

**S2**: A regular news reader is interested in reading different genres of news and faces the difficulty of capturing the content summary of a long news article effectively. While headlines and short descriptions of the news provide a general idea of the content, sometimes the information provided is insufficient. The situation worsens while reading on a mobile device as
the reading ability is limited by the screen dimension. This reader would also like to have the utility to bookmark the articles for later reading if he is in a rush or on the move.

S3: A basketball fan would like to follow up on the news articles about NBA but there is no such column from his preferred news providers and the sports section is flooded with news of other types of sports in which he is not interested. He would like to be able to filter news articles and receive updates on news related to basketball only.

S4: A weekly entertainment magazine journalist, who is also a keen entertainment news follower is concerned about the news in the media industry. She is assigned with an assignment on Britney Spears and also wants to gather more news about other celebrities that is related to Britney. She needs to mark articles that are useful for her and share them with other workmates within the company.

S5: A financial speculator relies on reading news articles to understand the trends and the movements of the market and to track how is it being affected by the decisions of key players and government policies. He requires good coverage on news articles from the entity that he selected and he would like to have some tools or functionality that help to organise the articles so that he can access them at a later time for visualisation, analysis and producing reports.

6.2 SYSTEM DETAILS

The system is programmed using ASP.NET 3.5 written in C# language and implemented on Windows Server 2003. SQL Server 2008 with full text indexing services enabled is used as the main storage to provide efficient text indexing and searching. Asynchronous JavaScript and XML (AJAX) scripting are used to provide a highly responsive webpage with minimal amount of page reload and server post back.

Figure 6.1 depicts the main Graphical User Interface (GUI) of our system. The components and layout are designed to support users to quickly filter the news headlines according to their interest. A tag cloud is displayed to indicate the top most important terms according to their category. Users can browse in both the traditional sectional view and entity filtering. The pictures panel displays interesting images from the news stream with their captions. Registered users can then login to the system to enjoy full features. The function of the components are summarised in Table 6.1.

News articles are categorised into high level sections by their nature, e.g: World News, Australia, Business, Entertainment, Sports, Health, Science and Odd. This browsing mode allows users to read the news based on their interests and aims to imitate the reading experience of reading a hardcopy newspaper. Headlines are displayed in the panel and sorted in reverse
chronological order to allow readers to discover latest news breaking stories. In addition, the system allows users to customise the order of the category presentation to show news according to their preference. By default only headlines of the same day will be displayed and that can be modified by selecting different modes (such as 7 days or 30 days) to include older headlines.

<table>
<thead>
<tr>
<th>Component</th>
<th>Name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Section bar</td>
<td>Allow user to toggle between sections</td>
</tr>
<tr>
<td>B</td>
<td>Mode selector</td>
<td>Allow user to filter news according to the timeframe</td>
</tr>
<tr>
<td>C</td>
<td>Login panel</td>
<td>Enable registered to login to the system</td>
</tr>
<tr>
<td>D</td>
<td>Quick filter</td>
<td>Filters for user to show related news to selected entities</td>
</tr>
<tr>
<td>E</td>
<td>Photos panel</td>
<td>To show interesting images from the news feed</td>
</tr>
<tr>
<td>F</td>
<td>Headlines panel</td>
<td>Display for latest news</td>
</tr>
<tr>
<td>G</td>
<td>Popular tags</td>
<td>Top 25 tags with most frequent appearance</td>
</tr>
</tbody>
</table>

Table 6.1. GUI Components

![Figure 6.1. Main Graphical User Interface](image)

6.3 FUNCTIONALITY

The OnestopNews system aims to provide innovative functionality which aims to extend present news systems:

© 2009 Cher Han (Andy) Lau
6.3.1 Reading tools

One of the basic functionalities that is lacking in current news browsing is the function to identify entities in the content. To address this limitation, we provide a set of extensive tools which help the users in reading and understanding the content of the news. For quick scanning, users can utilise the highlighter toolbar to identify NE’s by selecting the category or using the accordion panel that summarises entities in the news articles by their category or by their relevance. The importance of the entity is calculated based on the TF-IDF score. Users can identify the occurrence of the entity by clicking, and a popup box with the entity information from Wikipedia will appear upon hovering. Users can add the news to their clipbook and assign positive or negative votes to indicate their preference for the article. The highlighter toolbar allows users to highlight the selected entity. This set of reading tools aims to assist users to visually locate important entities and facilitate the reading process.

A key feature in online news which cannot be achieved by printed news is the ability to dynamically show related information. The related news panel provides users with a guide to link news articles together which helps readers to understand the development of an event as a whole. We further extend the concept of related news and implement related entity panels which show other entities that are closely related to the news article. Related entities will function as an indicator to signify the importance of another entity, which further strengthens readers understanding in conceptualising the overall evolution of a particular entity or visualising the connection between each entity.

![Figure 6.2. Related news and related entity panel](image)
Chapter 6: OneStopNews

Battered Wall Street gives to Obama, McCain

Some of John McCain’s and Barack Obama’s biggest donors are executives from the shaken financial services industry, which will need all the help it can get from whoever wins the White House.

Merrill Lynch & Co.’s chief executive, for example, has raised more than $500,000 for McCain’s campaign. Obama has received at least $1.5 million collected by three senior executives at Lehman Brothers.

McCain and Obama are both carefully courting the Wall Street money, Securities and investment firms gave $8.9 million to Obama and $5.9 million to McCain, according to data released by the Center for Responsive Politics.

The tough talk is coming from candidates who have fueled their campaigns with Wall Street money. Securities and investment firms gave $8.9 million to Obama and $5.9 million to McCain, according to data released by the Center for Responsive Politics.

Wall Street-based firms were among the most active in "bundling" contributions for the two candidates.

Three executives from Goldman Sachs Group Inc. have raised at least half a million dollars for Obama. That firm’s Obama’s top source of campaign money overall. Its employees have contributed more than $650,000 to his campaigns, according to the center.

Barack Hussein Obama II (born August 4, 1961) is the junior United States Senator from Illinois. He is the presumptive nominee of the Democratic Party in the 2008 presidential election, and the first African American to be a major party candidate for President of the United States. A graduate of Columbia University and Harvard Law School, Obama started as a community organizer and practiced as a civil rights attorney before serving in the Illinois Senate from 1997 to 2004. He also served as a constitutional law professor at the University of Chicago Law School from 1992 to 2004 following an unsuccessful bid for a seat in the U.S. House of Representatives in 2000, he announced his campaign for the U.S. Senate in January 2003. After winning a landslide primary victory in March 2004, Obama delivered the keynote address at the Democratic National Convention in July 2004. He was elected to the Senate in November 2004 with 70% of the vote. As a member of the Democratic minority in the 109th Congress, he promoted legislation to control conventional weapons and to promote greater public accountability in the use of federal funds. He also made official visits to Eastern Europe, the Middle East, and Africa. In the current 110th Congress, he has sponsored legislation regarding lobbying and election fraud, climate change, nuclear terrorism, and his defense of U.S. military personnel. Since announcing his presidential campaign in February 2007, Obama has emphasized withdrawing American troops from Iraq, increasing energy independence, decreasing the influence of lobbyists, and promoting universal health care as top national priorities.

© 2009 Cher Han (Andy Lau)
6.3.2 Support for named entity based searches

OneStopNews utilises the news indexing results to enable users to complement keyword based searches by using an NE. Users can choose from a list of possible meanings of a search term and conduct a search using the meaning of the term. The users can then decide the entities to be included and excluded. The filters can be saved for future use and the search results can be clipped into the clipbook.

---

![Figure 6.5. Search criteria formulation](image)

**Figure 6.5. Search criteria formulation**

![Figure 6.6. Search result for news article](image)

**Figure 6.6. Search result for news article**

![Figure 6.7 Search result for entity](image)

**Figure 6.7 Search result for entity**
6.3.3 Tagging

Tagging is implemented in the system as the main feature to assist users in organising information for its effectiveness of self-organisation [16] and the ability to improve social navigation [79]. Users can utilise the tagging functionality to complement automatic NE extraction, which provides a means to refine the content of the news article. A good scenario of such will be the entity: “Summer Olympics 2008,” where users obtain a reasonably decent coverage for all the articles about 2008 Olympic events. The reader can use the tags to assist them in further classifying the articles according to the nature of the events (such as Swimming, Archery, Taekwondo, etc). In this case, he can create the tag ‘olympic-taekwondo’ to tag news related to Taekwondo in the Olympics. In another example, there are a lot of news articles mentioning ‘Kevin Rudd’ and a user may only be interested in some but not all of the news. The user can simply utilise tags such as ‘rudd’ and assign the meaning of ‘Kevin Rudd’ to the tag to help them to mark off the news that they are interested in.

Figure 6.8 shows the tagging toolbar provided by the system to support the tagging operation. The dictionary is integrated to provide additional control of the tag semantics. Users can select any short descriptive label without any restriction on the choice of tag to apply to the article and the system will recognise the tag automatically and match the meaning of the tags in the dictionary. With the dictionary integration, users can now assign a meaning to the tag and therefore the same tags can be repeatedly used with different meanings without affecting the overall retrieval performance. If the user tag cannot be recognised by the system, the user can decide on the tags as their personal tags to indicate their personal usage, or they can choose another one of the meanings from the dictionary.

In order to promote social efficiency and achieve a high collaborative level, the system also provides contextual tag suggestions to users. Tag suggestion implementation provides users with a quick idea of what tags others are using and can effectively solve some problems such as cold start (i.e. do not know what are the appropriate tags to use) and typographical errors [83]. The tag suggestions will encourage users to reuse tags which will ultimately result in a more compact and converged tagging space.
6.3.4 Event Follow up

Manually following up development of specific news events can be tedious and laborious. The task can become more complicated if exploring a subset of news (e.g. to follow up on swimming news only in Olympics) or less popular news (i.e. news publishers do not create dedicated columns for the topic). News clipbook functionality is implemented in the system to provide support of article management, which allows users to create personalised interest topics.

Two types of clipbook operation: automatic and manual are available. In the automatic mode, users can create clipbooks by defining a start date and entities to be included and entities to be excluded. News articles can be picked up automatically and users will be informed whenever a new article is added. A manual clipbook allows users to add articles to the collection while reading. Users can define a global “exception list” in their personal settings to eliminate the certain named entities they are not interested. For example, users can define “NFL” to avoid reading news articles related to American Football League and “oil” for petrol related news. This helps the users by personalising the news articles presentation according to their own need and balancing the content if the articles are biased. Figure 6.9 shows an example of the clipbooks collection with the thumbnails automatically extracted from the news photo within the article.
6.3.5 Integrated social support

At the present moment news sharing functions provided by news systems are generally relying on external social websites without much moderation. Figure 1 models the activity between user interactions in a current news system. If user A needs to find the news articles of User B, C and D, he will need to be registered and login to all the bookmarking sites prior to being able to share and discover information about them. Social network systems provide a solid facility to help the users in managing relationships but the features available for the sharing mechanism are limited to sharing articles through the form of links submission. In addition, the absence of centralised control authority makes it impossible for users to track their activities and learn about information from other users.

OneStopNews implements a social component to facilitate users in performing social activities as shown in Figure 6.11. The social system manages both the user’s network and the activity log. Users can add other registered users as their friends and share information among the network. Users can keep track of their own activities and also learn about trends within the community. The aim of this is to promote interaction with other users. Every user will be provided with a personal page to display their activities and preferences.

![Figure 6.10. Current news sharing model](image-url)
6.3.6 Visualisation

The amount of information being delivered is increasing rapidly and therefore users require an effective method to support navigation and visualisation of the entities and tags. The visualisation scheme aims to help the user to identify the important entities that have been widely mentioned or are actively appearing in the news. We adopt a tag cloud as the visualisation tool in this project for its ability in representing important entities using a different font size and allows users to easily identify key terms efficiently. Another functionality provided by this system is the support of serendipity operations for users who aim to explore the news space to look for something to read and to gain information. A tag cloud is a collection of named entities listed with different font sizes to represent their weights. The weight is determined by the occurrence of the entities. For example, Figure 6.12 shows three examples of tag clouds. We can easily identify that the important persons are Barack Obama, John McCain and Sarah Palin, events such as Golden Globe, Grammy Award and Commonwealth Day. Users can also easily identify that there are a number of news items mentioning locations such as Boston, Cleveland and Washington DC. This presentation method helps the users to visualise key entities quickly and effectively.
Another important dimension in the social media is to capture and visualise items of interest within the community. In news websites and news aggregator websites, the importance of news is decided by the editor depending on their professional judgement. In a social context, the ‘interestingness’ can be reflected based on the voting results in vote-based recommendation systems such as Digg, Mixx and Reddit. A binary voting system is enabled in our system so that users can assign a positive vote to the stories that they like or give a negative vote for those that they do not like.

Users can rank the news according to the number of positive votes assigned by other users or friends to learn about what news is of interest within the community. This helps to explore the community news space and develop social activity formed around tags in a collaborative sharing environment. Use of tags represents a group of people or demonstrates a community language. Users that share similar levels of understanding use the same language and subsequent taggers can enrich the existing tags by providing more semantics to the resource. Therefore, users can retrieve news related to the topics that they are interested in or get recommendations by exploring the news that their friends are interested in.

![User Activity Log](image)

**Figure 6.13. User activity log**

### 6.4 CONCLUSION

In this chapter, we present OnestopNews, a collaborative news system with named entity indexing and social tagging enabled. The system provides additional functionality based on the following areas:
• **Reading and retrieval**: To utilise reading tools to locate, identify and understand the content of the article. Advanced search and entity based searching is provided to offer higher quality searches, which generate concise results with less noise and wider topic coverage.

• **Organisation and management**: Tools are provided to allow users to annotate articles using tagging, a clipbook to collect news articles and a log to review past activity.

• **Collaborative sharing**: The functionality is offered to allow users to share interesting articles, manage friends and discover and learn new knowledge from the trends of the whole community.

Reverse chronological news presentation order remains popular and useful as users tend to be interested in the most recent news. As shown by the printed newspaper model, today’s news is definitely of greater interest than yesterday news. However, readers are not interested in every news article. Keyword search functionality provided by most news websites is useful when users have some topics in mind. However certain users are unable to find news that is interesting to them if they do not have any specific topic in mind. The NE tags extracted automatically from the content help them to concisely identify the important entities that appear from the context based perspective. Users can visualise the needs from the user perspective by toggling. We also show that case tagging is a good way to organise content efficiently if it is managed properly.

News events develop rapidly and the value depreciates swiftly as a result of self-evolvement. Due to the screen dimension and dynamism of the news nature, headlines in this hour will be replaced by some other news automatically in the next hour. This may cause a user to miss out on important articles that might be interesting to them. While this gap is currently filled up by using tools such as RSS feed reader to act as a robot (such as Google Reader) to retrieve headlines periodically, it might cause a certain level of difficulty for readers to digest if the amount of headlines grows overly large. For example, a user who went on holidays for a few days will realise the accumulated headlines in the RSS Reader had already gone beyond the hundreds count. Our system aims to address this problem by allowing users to organise knowledge to create news clipping collections by including and excluding NE’s, this helps the user to monitor the breaking news stories more effectively. In addition, the users can browse through the bookmarks and news clippings created by friends which allows them to capture a more complete picture of the news evolvement within the social circle in an easier and more interesting way.
Online news should have a more effective indexing scheme for data management and presentation. We do not discount the fact that browsing headlines and topics is inefficient, but users will definitely benefit from having more flexibility in deciding the delivered content. The proposed system is a more complete and structured online news system, with dictionary integration, showing that users can perform effective news retrieval within their personal selection of topics and be able to share information among friends and gather information more easily.

The proposed system is not only simply another news browsing system with tagging functionality included. The system identifies various issues and limitations in the current news browsing system and intends to provide a solution to fix the gaps and develop a better system. The evaluation results demonstrated that integration of various components will definitely improve the overall performance of the system. Furthermore, with the controlled vocabulary dictionary in place, the quality of tags and the information retrieval performance has achieved a significant improvement.
Chapter 7: Results and Discussion

The primary goal of an information retrieval system is to achieve a high recall rate to ensure most results are covered together with a high precision rate to avoid returning irrelevant results. This chapter presents the evaluation of both the tag mapping algorithm and the effect of applying the dictionary as controlled vocabulary in a collaborative tagging news system.

7.1 SCOPE

The evaluation of this research is performed by two information retrieval experts. It is designed and conducted with the following aspects:

**Mapping accuracy**

This will measure the accuracy of tag mapping algorithm and validate the tags are successfully mapped to its designated node in the knowledge base. This is to ensure the users are presented with precise meaning of a named entity to facilitate the understanding the article content. In addition, the mapping performance will directly affect the tagging assistance hence plays an important role in overall retrieval performance. This also guarantees the correct entity is highlighted during reading and grouped into respective category during visualisation.

**Dictionary**

This evaluation demonstrates the ability of dictionary in gaining additional information from external sources and to retrieve related entity. A set of examples with various dictionary entries and the snapshot of related tags at different times will be presented which shows the change of related entity across a period.

**Tag convergence**

This section investigates the effect on the tagging space after the implementation of controlled vocabulary. The size of a tagging space and its growth rate directly affect the information retrieval performance.

**Retrieval enhancement**

This section provides a comparison of the retrieval performance between keyword based and named entity based search. The evaluation investigates whether named entity based retrieval with the dictionary will help to increase the articles coverage and filter off irrelevant articles and its efficiency.
User evaluation

This section evaluates the user satisfaction level towards the overall functionality provided by the system. The focus of the survey is trying to understand how the user can be beneficial from the various new features provided such as sharing with friends, important entity visualisation using tag cloud and the use of dictionary to assist them during tagging.

The metrics used in this evaluation are namely precision, recall and F-score. Precision measures the relevancy by representing a ratio of correctly retrieved items and the total items retrieved but does not show whether all relevant items are retrieved. Recall represents the system capability in retrieving the relevant items by representing a ratio of correctly retrieved item over the total items retrieved, but the amount of irrelevant item is not represented. The two measurements are then combined into an F-score to represent the weighted harmonic mean of precision of recall as an indicator of overall performance.

7.2 DATASET

News data of all category (e.g. World, Local, Entertainment, Sports, Technology, Health and Science) from ABC and Newsvine (from 26/08/2008 to 29/08/2008) are selected as the primary dataset for use in this research project. ABC news provides tags for users as a secondary category to follow news articles and Newsvine allows content submitters to assign tags to describe articles. The entire dataset contains 5000 news articles with 32565 named entity tags, 16427 creator tags and 12192 user tags.

7.3 DICTIONARY

A dictionary is constructed automatically to show the tags mapping result. The dictionary is presented in Resource Description Framework (RDF) and can be consumed as controlled vocabulary or connected to other systems for data enrichment.

Figure 7.1 illustrates the RDF example of the entity instance “Australia.” On top of the property that exists in the original document, the model is enriched by connecting to DBPedia [138]. DBPedia knowledge base extracts structured information from Wikipedia and provides an interface that allows external systems to query the system. For instance, additional information about a country such as geographical information, names in other languages, capital and leader’s names can be extracted. This demonstrates the capability of interoperability with other systems which is based on the same unique identifier. As both the dictionary and DBPedia are based on Wikipedia, the data model can be exchanged and enriched easily. For instance, the enrichment model demonstrates that the entity ‘Australia’ can be enriched with additional information such
as geographical information (capital, longitude and latitude), leader information and name translations in other languages.

![Graphical representation of 'Australia' RDF after enrichment](image)

Figure 7.1. Graphical representation of 'Australia' RDF after enrichment

Table 7.1 reveals the statistics of another two instances of the dictionary entry Lehman Brothers and Kevin Rudd. Lehman Brothers represents a popular entity that appears in 23 different writing forms in the dataset and Kevin Rudd appears with less variation. The system is able to detect the relevant entity accurately. As a unique identifier is selected, rich amounts of information can be further enriched from the DBPedia system. Another unique feature enabled by the dictionary is the ability to list related entities provided when being queried at different times. Figure 7.2 shows the snapshot of the related entities at different times for “Kevin Rudd” which enables users to understand related entities easily.

<table>
<thead>
<tr>
<th>Dictionary Entry</th>
<th>Total surface forms</th>
<th>Average related entity per article</th>
<th>DBPedia enriched property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lehman Brothers</td>
<td>23</td>
<td>5.977</td>
<td>27</td>
</tr>
<tr>
<td>Kevin Rudd</td>
<td>6</td>
<td>4.471</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 7.1. Statistics of two dictionary instances
7.4 MAPPING ACCURACY

The success of the overall system performance relies on the accuracy that the mapping algorithm provided. These results aim to report the performance of the overall tag mapping algorithm and its individual performance while being applied to different tags categories. The robustness of the mapping algorithm is measured by the following:

\[
\text{Precision} = \frac{\text{correctly mapped tag } \cap \text{ retrieved tag}}{\text{correctly map tag}} \tag{4}
\]

\[
\text{Recall} = \frac{\text{correctly mapped tag } \cap \text{ retrieved tag}}{\text{retrieved tag}} \tag{5}
\]

\[
F = \frac{2 \times \text{Precision } \times \text{Recall}}{\text{Precision } + \text{Recall}} \tag{6}
\]

A NE tag is considered to have mapped correctly if it is mapped to the correct Wikipedia identifier or the closest related concept if a wiki page is absent. Categorical tags such as science, economics and health are not covered in the scope of evaluation as they are automatically categorised into the relevant news section. Tags that failed to be mapped will be manually inspected by the expert and will not be included in the evaluation. The examples of such are shown below.

<table>
<thead>
<tr>
<th>Australia</th>
<th>Australia</th>
<th>Malcom Turnbull</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malcom Turnbull</td>
<td>United States</td>
<td>Pakistan</td>
</tr>
<tr>
<td>Robert McClelland</td>
<td>Barack Obama</td>
<td>United Nation</td>
</tr>
<tr>
<td>Brendan Nelson</td>
<td>Tasmania</td>
<td>Brendan Nelson</td>
</tr>
<tr>
<td>Federal Government</td>
<td>Malcom Turnbull</td>
<td>Chris Bowen</td>
</tr>
<tr>
<td>Melbourne</td>
<td>Federal Government</td>
<td>George Bush</td>
</tr>
<tr>
<td>United States</td>
<td>Wayne Swan</td>
<td>Hamid Karzai</td>
</tr>
<tr>
<td>Wayne Swan</td>
<td>Israel</td>
<td>Isalambad</td>
</tr>
<tr>
<td>Rob Stary</td>
<td>George Bush</td>
<td>Ban ki-moon</td>
</tr>
<tr>
<td>Liberal Party</td>
<td>China</td>
<td>Reserve Bank of Australia</td>
</tr>
</tbody>
</table>

(a) 08/09/08 to 14/09/08  (b) 15/09/08 to 21/09/08  (c) 10/11/2008 to 16/11/2008

Figure 7.2. Related entities of “Kevin Rudd” at different times
### Table 7.2. Example of correctly mapped tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Wikipedia identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle-restaurant</td>
<td>Seattle, Washington</td>
<td>Restaurant information in Seattle</td>
</tr>
<tr>
<td>Google-phone</td>
<td>Google Android</td>
<td>Mobile phone developed by Google</td>
</tr>
<tr>
<td>Kung-fu</td>
<td>Chinese Martial Arts</td>
<td>Chinese martial arts</td>
</tr>
<tr>
<td>allie-finkle</td>
<td>Meg Cabot</td>
<td>A comic drawn by Meg Cabot</td>
</tr>
<tr>
<td>Craig Tiley</td>
<td>2008_Australian_Open</td>
<td>One of Australian Open 2008 directors</td>
</tr>
</tbody>
</table>

### Table 7.3 Example of incorrectly mapped tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Wikipedia identifier</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>chinese-dairy</td>
<td>Chinese cuisine</td>
<td>2008_Chinese_milk_scandal</td>
</tr>
<tr>
<td>Cvs</td>
<td>Concurrent Versions System</td>
<td>CVS_Corporation</td>
</tr>
<tr>
<td>ice-age-man</td>
<td>Link</td>
<td>Ice_Age</td>
</tr>
<tr>
<td>lsu-vs-auburn</td>
<td>2006 LSU Tigers football team</td>
<td>Auburn_LSU_rivalry</td>
</tr>
<tr>
<td>Mac</td>
<td>MAC address</td>
<td>Macintosh</td>
</tr>
</tbody>
</table>

### Table 7.4. Named entity tag result

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>90.00%</td>
<td>92.86%</td>
<td>0.9140</td>
</tr>
<tr>
<td>Location</td>
<td>95.43%</td>
<td>97.56%</td>
<td>0.9648</td>
</tr>
<tr>
<td>Organization</td>
<td>95.68%</td>
<td>96.23%</td>
<td>0.9595</td>
</tr>
<tr>
<td>Event</td>
<td>95.67%</td>
<td>94.78%</td>
<td>0.9523</td>
</tr>
<tr>
<td>Others</td>
<td>94.25%</td>
<td>93.25%</td>
<td>0.9384</td>
</tr>
<tr>
<td>Average</td>
<td>95.56%</td>
<td>96.01%</td>
<td>0.9580</td>
</tr>
<tr>
<td>Category</td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>Person</td>
<td>96.78%</td>
<td>98.85%</td>
<td>0.9750</td>
</tr>
<tr>
<td>Location</td>
<td>92.31%</td>
<td>98.86%</td>
<td>0.9547</td>
</tr>
<tr>
<td>Organization</td>
<td>94.74%</td>
<td>95.12%</td>
<td>0.9492</td>
</tr>
<tr>
<td>Event</td>
<td>92.30%</td>
<td>91.12%</td>
<td>0.9170</td>
</tr>
<tr>
<td>Others</td>
<td>94.70%</td>
<td>91.03%</td>
<td>0.9280</td>
</tr>
<tr>
<td>Average</td>
<td>92.81%</td>
<td>93.80%</td>
<td>0.9326</td>
</tr>
</tbody>
</table>

Table 7.5. Creator tag result

<table>
<thead>
<tr>
<th>Tags Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>88.15%</td>
<td>97.56%</td>
<td>0.9171</td>
</tr>
<tr>
<td>Location</td>
<td>100%</td>
<td>98.92%</td>
<td>0.9945</td>
</tr>
<tr>
<td>Organization</td>
<td>96.00%</td>
<td>92.59%</td>
<td>0.9426</td>
</tr>
<tr>
<td>Event</td>
<td>96.56%</td>
<td>94.12%</td>
<td>0.9532</td>
</tr>
<tr>
<td>Others</td>
<td>92.86%</td>
<td>91.43%</td>
<td>0.9213</td>
</tr>
<tr>
<td>Average</td>
<td>94.71%</td>
<td>94.92%</td>
<td>0.9457</td>
</tr>
</tbody>
</table>

Table 7.6. User tag result

Figure 7.3. Mapping performance result
The algorithm produces excellent performances with the majority of the tags being correctly identified and mapped. NE tags produce the best results expectedly as the entities are extracted automatically from the source content. Location category in user tags achieved the best performance while the Person category in NE is the worst. This is due to the lesser variations in the nature of a location tag’s surface form. Well-known NE tags such as “Australia”, “Indonesia” and “United Kingdom” are identified and recognised successfully. In addition, unpopular NE tags “Lizard Island”, “Hefei” and “Gansu” are also successfully mapped. Unpopular tags such as “Suleymaniyah” is not recognised and “Pando” is identified as others due to their low occurrence usage. Location categories in user tags achieved the best performance since most of the users only tag the articles with well-known places.

A similar observation is made on the user tag in the Person category. Examples of person tags that are unable to be retrieved can be non-famous sports players, local officers and spokespersons of an organisation. These tags failed to be mapped due to the following reasons: (i) their name appeared at multiple different Wikipedia pages (but they do not own a Wikipedia page) and (ii) Their related Wikipedia pages are not within the Top 20 expert results. High popularity entity tags such as “Barack Obama” and their related topics (e.g. Obama-camp) and surface forms (e.g. obama, barack-obama) are covered effectively.

The mapping algorithm produces satisfactory results on Organisations, Events and Others categories as the variations between terms are less significant and easily detectable. The high usage term “Lehman Brothers” is one of the terms with most variations, however all of the surface forms contain the word “Lehman” (i.e. lehman-brothers, Lehman Brothers Holdings Inc., Lehmans) which makes the task less difficult. The algorithm is also capable of processing complex tags such as ‘the anniversary of the 2001 terrorist attacks’ and ‘the Sept 11 anniversary’ and connect to the KB identifiers ‘September 11, 2001 attacks’. The significant case where the system is affected by confusion cases will be misinterpreting ‘SP 500’ as the chemical element Sulphur instead of ‘S&P 500 stock market’.

7.5 TAG CONVERGENCE

A key outcome that is expected from the system is the tagging space convergence. Tags distribution is sparse and generally follows Zipf’s Law (i.e. popular terms are used very frequently while large numbers of tags are rarely used) while the distribution is more even in traditional classification. The problem worsens when being applied in folksonomy due to the uncontrolled nature of tagging which causes the vocabulary space to be chaotic. For a social bookmarking system to be effective and collaborative, it is important to maintain the tag convergence to ensure the tags are reused appropriately and converge with the tagging space to maintain a reasonable vocabulary growth with the minimal amount of repetition.
In this section, we examine the effect of the controlled dictionary implementation and how it helps to achieve effective tag convergence in the system. Tag convergence represents the system's ability to group semantically similar or identical tags together to provide a higher recall rate and result in a compact tag space. Tag convergence will result in a tagging space with the least tags variations and present a more accurate popularity of the tag by reporting the actual usage of a term.

Table 7.7 shows the results of five representative entities that demonstrate the system's ability in capturing the root form and their surface forms and Table 7.8 shows the detailed categorical tag convergence efficiency after the controlled vocabulary implementation. Figure 7.4 indicates that the growth of the tag space is slower after the dictionary implementation. The results show that our system achieves a good performance by successfully reducing the tag space size by 30% as 943 surface forms of the entity are effectively grouped to their root forms. Similarly, Figure 7.5 illustrates the effect of tag convergence by their category. In the best case of event category the system manages to reduce 39% of the tag space and still achieve a minimum of 20% reduction for person category (as there are less variations of person names used in news articles).
<table>
<thead>
<tr>
<th>Root Form</th>
<th>Surface Forms</th>
<th>Occurrence %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama (Person)</td>
<td>Barack Obama</td>
<td>50.37%</td>
</tr>
<tr>
<td></td>
<td>Obama</td>
<td>25.50%</td>
</tr>
<tr>
<td></td>
<td>barack-obama</td>
<td>18.81%</td>
</tr>
<tr>
<td></td>
<td>democrat-barack-obama</td>
<td>2.48%</td>
</tr>
<tr>
<td></td>
<td>barack-obama-news</td>
<td>0.50%</td>
</tr>
<tr>
<td></td>
<td>Obama camp</td>
<td>0.50%</td>
</tr>
<tr>
<td></td>
<td>Barack</td>
<td>0.37%</td>
</tr>
<tr>
<td></td>
<td>(others)</td>
<td>1.49%</td>
</tr>
<tr>
<td>United Kingdom (Location)</td>
<td>United Kingdom</td>
<td>87.56%</td>
</tr>
<tr>
<td></td>
<td>united-kingdom</td>
<td>6.10%</td>
</tr>
<tr>
<td></td>
<td>Uk</td>
<td>3.76%</td>
</tr>
<tr>
<td></td>
<td>(others)</td>
<td>2.58%</td>
</tr>
<tr>
<td>Lehman Brothers (Organisation)</td>
<td>Lehman Brothers</td>
<td>10.77%</td>
</tr>
<tr>
<td></td>
<td>Lehman Brothers Holdings Inc.</td>
<td>3.93%</td>
</tr>
<tr>
<td></td>
<td>lehman-brothers</td>
<td>3.59%</td>
</tr>
<tr>
<td></td>
<td>Lehman</td>
<td>3.16%</td>
</tr>
<tr>
<td></td>
<td>Joe Bel Bruno</td>
<td>1.79%</td>
</tr>
<tr>
<td></td>
<td>Lehman Brothers Holdings Inc</td>
<td>0.68%</td>
</tr>
<tr>
<td></td>
<td>lehman-brothers-holdings</td>
<td>0.51%</td>
</tr>
<tr>
<td></td>
<td>Lehman Bros</td>
<td>0.26%</td>
</tr>
<tr>
<td></td>
<td>(others)</td>
<td>2.31%</td>
</tr>
<tr>
<td>Emmy Award (Event)</td>
<td>Emmys</td>
<td>36.47%</td>
</tr>
<tr>
<td></td>
<td>Emmy</td>
<td>33.53%</td>
</tr>
<tr>
<td></td>
<td>Emmy Awards</td>
<td>16.47%</td>
</tr>
<tr>
<td></td>
<td>emmy-awards</td>
<td>11.18%</td>
</tr>
<tr>
<td></td>
<td>Emmy award</td>
<td>2.35%</td>
</tr>
<tr>
<td>Bail_out (finance)</td>
<td>Bailout</td>
<td>86.44%</td>
</tr>
<tr>
<td></td>
<td>bail-out</td>
<td>5.08%</td>
</tr>
<tr>
<td></td>
<td>Bailouts</td>
<td>5.08%</td>
</tr>
<tr>
<td></td>
<td>(others)</td>
<td>3.39%</td>
</tr>
</tbody>
</table>

Table 7.7. Tag convergence performance
Table 7.8. Tag convergence performance (by category)

<table>
<thead>
<tr>
<th>Category</th>
<th>No of Tags</th>
<th>Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Person</td>
<td>1553</td>
<td>1265</td>
</tr>
<tr>
<td>Location</td>
<td>179</td>
<td>120</td>
</tr>
<tr>
<td>Organization</td>
<td>982</td>
<td>719</td>
</tr>
<tr>
<td>Event</td>
<td>58</td>
<td>35</td>
</tr>
<tr>
<td>Others</td>
<td>1239</td>
<td>929</td>
</tr>
<tr>
<td>Total</td>
<td>4011</td>
<td>3068</td>
</tr>
</tbody>
</table>

7.6 RETRIEVAL PERFORMANCE

The purpose of this evaluation is to verify the difference of retrieval performance between keywords and named entities. Ten topics were selected from the subset of the dataset to form a 5000 news articles subset ground truth for the evaluation using equation 5, 6 and 7.

Table 7.9 represents the evaluation result. Both precision and recall are over 80% with significant improvement ranged from 0.110 (US Election) to 0.307 (Olympic). The result shows that named entity based retrieval complements keyword based retrieval. The increment in recall rate shows that the system can retrieve relevant articles with terms in different spelling but with the same meaning. The improvement in precision rate indicates that the returned results contain much less noise as results that contain ambiguous terms with different meanings are eliminated.
<table>
<thead>
<tr>
<th>No</th>
<th>Keyword</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lehman Brothers</td>
<td>0.880</td>
<td>0.550</td>
<td>0.677</td>
<td>0.947</td>
<td>0.900</td>
<td>0.923</td>
</tr>
<tr>
<td>2</td>
<td>Barack Obama</td>
<td>0.889</td>
<td>0.686</td>
<td>0.774</td>
<td>0.914</td>
<td>0.970</td>
<td>0.941</td>
</tr>
<tr>
<td>3</td>
<td>Emmy Award</td>
<td>1.000</td>
<td>0.667</td>
<td>0.800</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>US Election</td>
<td>0.882</td>
<td>0.600</td>
<td>0.714</td>
<td>0.840</td>
<td>0.808</td>
<td>0.824</td>
</tr>
<tr>
<td>5</td>
<td>Olympic</td>
<td>0.850</td>
<td>0.486</td>
<td>0.618</td>
<td>0.912</td>
<td>0.939</td>
<td>0.925</td>
</tr>
<tr>
<td>6</td>
<td>NFL</td>
<td>0.868</td>
<td>0.596</td>
<td>0.707</td>
<td>0.927</td>
<td>0.934</td>
<td>0.930</td>
</tr>
<tr>
<td>7</td>
<td>Al Qaeda</td>
<td>0.892</td>
<td>0.649</td>
<td>0.751</td>
<td>0.932</td>
<td>0.904</td>
<td>0.918</td>
</tr>
<tr>
<td>8</td>
<td>New York</td>
<td>0.873</td>
<td>0.760</td>
<td>0.813</td>
<td>0.938</td>
<td>0.944</td>
<td>0.941</td>
</tr>
<tr>
<td>9</td>
<td>Osama Bin Laden</td>
<td>0.843</td>
<td>0.583</td>
<td>0.690</td>
<td>0.892</td>
<td>0.952</td>
<td>0.921</td>
</tr>
<tr>
<td>10</td>
<td>Hillary Clinton</td>
<td>0.832</td>
<td>0.639</td>
<td>0.723</td>
<td>0.927</td>
<td>0.964</td>
<td>0.945</td>
</tr>
</tbody>
</table>

Table 7.9. Information retrieval performance evaluation

7.7 USERS STUDY

A preliminary user survey has been conducted with 20 participants to verify the usability and effectiveness of the system. To quantise users’ input, we have used closed questions that can be answered with a 1-5 Likert scale ratings (1: Strongly agree, 2: Agree and 3: Neutral, 4: Disagree, 5: Strongly disagree). The following questions are used for evaluation.

Q1. The dictionary helps in improving results
Q2. The tag cloud of categories is useful for understanding key entities
Q3. The reading tools provided are useful
Q4. The tagging function helps in organising and retrieving more news
Q5. The news clipping function helps to follow up on breaking stories
Q6. The sharing options provided are sufficient

The individual users rating results can be found on  and the average rating for each question is depicted in . It is important to note that all users find the dictionary and tag cloud are useful for them. This agrees with our hypothesis that the lightweight indexing scheme is sufficient to control and manage the tags, which helps to enhance the result and indicates that a tag cloud is a good visualisation method. Furthermore, majority of the users find that the reading tools provide good assistance while reading news.
In Q1, 80% of the users find that the retrieval results are effective and meet their requirements. Users who feel neutral towards the question are generally satisfy with the search result as they either use a more restrictive keyword (e.g. “Barack Obama”) or they do not mind having presented news that are slightly irrelevant. One user who disagrees with this functionality thinks that selecting word sense appear to be an extra step for him during the search activity.

In Q2, none of the users disagree with the visualisation of the key entity using tag cloud. Most of the users who find it exceptionally useful and strongly agree with its usability indicate that it was their first time seeing such visualisation. In Q3, users who find the NE panel and reading tools such as highlighters are not useful indicate that they prefer to read the article from the beginning without skimming, but the majority of the users agree that it helps them to pay attention to the key person or location in the article.

In Q4 and Q5, users who indicate that tagging is not really helpful for them in organising the news articles at the same time points out that news clipping function is sufficient for them to track the evolution of news events. Interestingly, other users who find tagging but not news clipping useful, indicate that clipping functions are secondary as they have previous experience in other similar social website, therefore they are more comfortable in using tags to bookmark favourite articles and possibly share with other friends. Users who disagree with the tagging functionality point out that they are not convinced with tagging, as it appears to be unmanageable when the tags space grows big.

Lastly, majority of the users responded neutral to Q6 as they primarily use the news system for browsing purpose, which makes sharing options appear to be secondary. For users who disagree with this question, their comments are the system should also include email functionality in addition to the built-in sharing and facebook posting options. However, our main design consideration is to purposely exclude the email functionality so that the users will be focus only on evaluating the sharing feature.
Table 7.10 Summary of user study data

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Response Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The dictionary helps in improving retrieval result</td>
<td>11 (55%)</td>
<td>5 (25%)</td>
<td>3 (15%)</td>
<td>1 (5%)</td>
<td>0 (0%)</td>
<td>1.70</td>
</tr>
<tr>
<td>2. The tag cloud is useful for understanding key entities</td>
<td>10 (50%)</td>
<td>7 (35%)</td>
<td>3 (15%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1.65</td>
</tr>
<tr>
<td>3. The reading tools provided are useful</td>
<td>7 (35%)</td>
<td>7 (35%)</td>
<td>4 (20%)</td>
<td>1 (5%)</td>
<td>1 (5%)</td>
<td>2.10</td>
</tr>
<tr>
<td>4. The tagging function helps in organising and retrieving more news</td>
<td>3 (15%)</td>
<td>12 (60%)</td>
<td>3 (15%)</td>
<td>0 (15%)</td>
<td>2 (10%)</td>
<td>2.30</td>
</tr>
<tr>
<td>5. The news clipping functions helps in following breaking news stories</td>
<td>2 (10%)</td>
<td>9 (45%)</td>
<td>7 (35%)</td>
<td>1 (5%)</td>
<td>1 (5%)</td>
<td>2.50</td>
</tr>
<tr>
<td>6. The sharing options provided are sufficient</td>
<td>3 (15%)</td>
<td>5 (25%)</td>
<td>8 (40%)</td>
<td>2 (10%)</td>
<td>2 (10%)</td>
<td>2.75</td>
</tr>
</tbody>
</table>

7.8 DISCUSSION

This chapter presents various results generated from the system and the evaluation of the system performance. The aspect of the study focuses on the named entity indexing and its application by evaluating the mapping accuracy and the aftermath effect applied to the tagging. The results show the dictionary production and the mapping accuracy represents the main judging criterion which leads to the success of the whole system.

Section 7.5 shows the improvements in information retrieval after the integration of the NE dictionary as the controlled vocabulary. The information retrieval performance improved significantly by using the named entities which effectively retrieved more relevant results and less irrelevant results. The minimum F-score is 0.840 for Topic 4 (US election) due to wide coverage deviation of the topics. Topic 3 (Emmy Award) achieved the best precision performance for both keywords as the terms have less variation, but an NE search can effectively cover all the related articles.

The preliminary user study shows that the system has achieved a successful implementation. The tagging functionality extends the existing usability and provides users with additional features in organising articles and supporting serendipitous news discovery from personal preference or within the social context. The high accuracy named entity extraction,
together with tags suggestion and tag mapping, has formed a more concise and compact tagging vocabulary space which effectively addresses various social tagging deficiencies and the performance of information retrieval is greatly enhanced.
Chapter 8: Conclusion and Future Work

8.1 SUMMARY OF ACHIEVEMENTS

Collaborative tagging and folksonomy represents the main technology and plays an important role in driving the development of Web 2.0. Structuring tagging in the social tagging system represents an important work in folksonomy and forms the basis of this research. An expert based tags mapping algorithm is used to design the technique and the ability is verified using an online news system.

A named entity (NE) tags mapping algorithm structure has been developed by using the expert suggestion based model. We leverage on the mature search engine technology and exploit the facility provided by Wikipedia to uniquely map NE tags to their respective meaning in the knowledge base. The outcome is promising as the algorithm effectively groups similar terms with different surface forms together and is able to differentiate different meanings of the tag. The evaluation results show that the proposed model is robust and effective in identifying senses and can be used as an effective tool for understanding the news content semantically.

A dictionary that represents the output from the tag mapping results has been developed to model the tag mapping results. The design of the dictionary is adhered to the NISO controlled vocabulary guidelines and described using Dublin Core (DC) metadata. It uses Resource Description Framework (RDF) which is a widely recognised standard for data interchange, allowing information exchange with other established systems. The dictionary integration in the online news system further demonstrates that such a dictionary can be applied effectively for use as a controlled vocabulary to govern various different types of named entity based tags in a collaborative environment.

A collaborative news tagging framework has been introduced to demonstrate the efficiency of the tag mapping and its usefulness in improving the information retrieval result. The system does not only provide various functionalities that support different needs of different news readers, but it has also included advanced features to demonstrate the impact of setting up a controlled vocabulary and its usefulness in assisting users in organising the content and filtering off content.

8.2 EVALUATION PERFORMANCE

The success of an information retrieval system is ultimately determined by its ability to support users demands and application requirements. The entity tag mapping achieves this goal.
Chapter 8: Conclusion and Future Work

by providing the users with a better understanding for the content of a news article without having to go through the whole article but is able to estimate the content nature by using important named entities that appear in the article. In addition, a social news system has to be considered that will help to assist users in identifying the top news stories, organising them effectively, filtering unwanted news, sharing with friends easily and retrieving news efficiently in the future. Three aspects have been evaluated to show the robustness of the system:-

- **The accuracy of the tag mapping**, which determines the ability to map the named entity and explicitly identify their meanings. By using Calais to extract named entities together with Wikipedia for identifying the meaning, the system performed well with high detection and accuracy. The mapping algorithm has been tested against a wide range of news genres, each with unique characteristics. The three main types of named entity tags are; named entity tags, creator tags and user tags. Tags are further categorised based on their nature of use in the news article according to the categories of person, location, organisation, events and others. The system demonstrates the desired performance level using Recall Rate (RR) and Precision (Rate). RR reflects the rate of misdetection and PR reflects the rate of incorrect detection. In summary, the system achieved an overall RR of 96.01% and PR of 95.56%. The system performed the best while performing Location detection with 97.56% recall and 95.43% precision. In the worst case of the Person category, due to its high variability, the system still managed to maintain the performance up to 92.86% RR and 90.00% PR.

- **The effectiveness of the collaborative news tagging system** is the second aspect that was evaluated. With the control vocabulary in place, the system effectively bridges the gap between tags and semantic knowledge bases. In the retrieval evaluation, the system performance improved significantly by retrieving more related articles with less noise. Having the ability to identify and map the tags to the correct sense, the system demonstrates the ability in differentiating the tags and grouping semantically identical tags together to converge the tag space.

- **The third aspect is the user satisfaction** of the collaborative news tagging system. Users are expected to leverage on the efficiency of the controlled vocabulary, to be benefited by the improved system. The controlled vocabulary facilitates the users tagging process by providing senses and suggestions, and also complements the keyword search by enabling named entity based searches. The results from a user study of 20 participants further confirmed this effectiveness. The dictionary results in users being able to retrieve more related news, more interesting news and share...
news among friends. These activities are expected to be performed with the least amount of effort or intervention without changing previous browsing habits. Furthermore, users are expected to utilise the system in identifying the news that they are interested in more effectively and have a quick overview of the news that is happening based on the various visualisation tools provided.

8.3 TRENDS AND FUTURE WORK

While active debates between folksonomy and ontology are going on, social networking has undoubtedly become one of the next generation online tools for users to organise and share content. Alterations and enhancements that improve the quality of tag efficiency overcome the fundamental problems of free form tagging and provide better retrieval performance. This will definitely be beneficial for the users and eventually attract more users to utilise tagging.

The dictionary is currently presented in the Resource Description Framework (RDF) format. In the future, we aim to enable the system to allow news articles using SPARQL language. In addition, it should be connected to other domain-specific knowledge bases to obtain more information.

While reading news is the formal channel for users to obtain information, the arising of blogs and wikis cannot be neglected as they represent another major channel for users to learn and share information within the blogosphere. In the future, we aim to extend the tag mapping algorithm for use in blogging and wiki writing platforms and provide strong interconnectivity. This will facilitate news readers in gathering current affairs information and at the same time discover other relevant information from within their social network.

At the moment, RSS news reader helps the news feed subscriber to robotically collect news headlines so that they will not miss out any headlines, but this does not mean that every headline is interesting or important to them. Digg and similar systems provide a platform for users to share interesting information and monitor friends activities easily but content organisation, filtering and future retrieval remain a mystery. Delicious and Newsvine provides tagging functionality which helps users to organise the content using tags, however, the tags are unmanaged which causes the overall system performance to drop significantly.

While the preliminary user study shed light on the research directions which agrees that such indexing scheme is lightweight but also effective to provide additional features which improves the users experience from retrieval and organising perspective. However, while the results appear to be promising, the evaluations should be continuously expanding by using a larger data corpus and more participants from different user groups.
In conclusion, our system represents a system that utilises multiple information sources ranging from both news and social information. Future work will definitely be conducted in this direction to facilitate structured tagging retrieval to provide users with a user-friendly environment for sharing and retrieving multimedia information efficiently via multiple channels.
Bibliography


[131] “RDF/XML Syntax Specification (Revised).”


