

---

# **A Generic Architecture for Semantic Enhanced Tagging Systems**

---

**PhD Thesis**

**Murad Magableh**

This thesis is submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy

---

Software Technology Research Laboratory  
De Montfort University  
Leicester - United Kingdom

*July 2011*

# Dedication

To our **Prophet Mohammed** (peace be upon him) for he said:  
“Whoever seeks a way to acquire knowledge Allah will make easy  
his way to paradise”.

To my **Mother** for her endless love, support, encouragement, and  
continuum prayers.

To my **Father's Soul** who was strong, faithful, and true. Wish you  
were here.

To my **Family**; sisters, brothers, nieces, and nephews for their love.

# Abstract

The *Social Web*, or Web 2.0, has recently gained popularity because of its low cost and ease of use. Social tagging sites (e.g. Flickr and YouTube) offer new principles for end-users to publish and classify their content (data). Tagging systems contain free-keywords (tags) generated by end-users to annotate and categorise data. Lack of semantics is the main drawback in social tagging due to the use of unstructured vocabulary. Therefore, tagging systems suffer from shortcomings such as low precision, lack of collocation, synonymy, multilinguality, and use of shorthands. Consequently, relevant contents are not visible, and thus not retrievable while searching in tag-based systems.

On the other hand, the *Semantic Web*, so-called Web 3.0, provides a rich semantic infrastructure. Ontologies are the key enabling technology for the Semantic Web. Ontologies can be integrated with the Social Web to overcome the lack of semantics in tagging systems.

In the work presented in this thesis, we build an architecture to address a number of tagging systems drawbacks. In particular, we make use of the controlled vocabularies presented by ontologies to improve the information retrieval in tag-based systems. Based on the tags provided by the end-users, we introduce the idea of adding “*system tags*” from semantic, as well as social, resources. The “*system tags*” are comprehensive and wide-ranging in comparison with the limited “*user tags*”. The system tags are used to fill the gap between the user tags and the search terms used for searching in the tag-based systems. We restricted the scope of our work to tackle the following tagging systems shortcomings:

1. The lack of semantic relations between user tags and search terms (e.g. synonymy, hypernymy),
2. The lack of translation mediums between user tags and search terms (multilinguality),
3. The lack of context to define the emergent shorthand writing user tags.

---

To address the *first* shortcoming, we use the WordNet ontology as a semantic lingual resource from where system tags are extracted. For the *second* shortcoming, we use the MultiWordNet ontology to recognise the cross-languages linkages between different languages. Finally, to address the *third* shortcoming, we use tag clusters that are obtained from the Social Web to create a context for defining the meaning of shorthand writing tags.

A prototype for our architecture was implemented. In the prototype system, we built our own database to host videos that we imported from real tag-based system (YouTube). The user tags associated with these videos were also imported and stored in the database. For each user tag, our algorithm adds a number of system tags that came from either semantic ontologies (WordNet or MultiWordNet), or from tag clusters that are imported from the Flickr website. Therefore, each system tag added to annotate the imported videos has a relationship with one of the user tags on that video. The relationship might be one of the following: synonymy, hypernymy, similar term, related term, translation, or clustering relation.

To evaluate the suitability of our proposed system tags, we developed an online environment where participants submit search terms and retrieve two groups of videos to be evaluated. Each group is produced from one distinct type of tags; user tags or system tags. The videos in the two groups are produced from the same database and are evaluated by the same participants in order to have a consistent and reliable evaluation. Since the user tags are used nowadays for searching the real tag-based systems, we consider its efficiency as a criterion (reference) to which we compare the efficiency of the new system tags.

In order to compare the relevancy between the search terms and each group of retrieved videos, we carried out a statistical approach. According to Wilcoxon Signed-Rank test, there was no significant difference between using either system tags or user tags. The findings revealed that the use of the system tags in the search is as efficient as the use of the user tags; both types of tags produce different results, but at the same level of relevance to the submitted search terms.

# **Declaration**

I declare that the work described in this thesis is original work undertaken by me for the degree of Doctor of Philosophy, at the software Technology Research Laboratory (STRL), at De Montfort University, United Kingdom.

No part of the material described in this thesis has been submitted for any award of any other degree or qualification in this or any other university or college of advanced education.

This thesis is written by me and produced using  $\text{\LaTeX}$ .

**Murad Magableh**

## **Publications**

- Murad Magableh, Antonio Cau, Hussein Zedan, and Martin Ward. Towards a multilingual semantic folksonomy. In *Proceedings of the IADIS International Conferences Collaborative Technologies 2010 and Web Based Communities 2010*, pages 178-182, July 2010.

# Acknowledgement

First and foremost, I humbly thank Allah who gave me health, thoughts, and cooperative people to enable me achieving this goal.

My supervisory team has been behind the achievement of this work. Their sage advices, wide knowledge, unfailing patience, and kind collaboration from the first day enabled this work to take place today. Thank you **Prof. Hussein Zedan, Dr. Antonio Cau,** and **Dr. Martin Ward.**

I am indebted to acknowledge and extend my genuine gratitude to **my teachers** all the way through my study period; primary school, secondary school, undergraduate, and postgraduate.

A special appreciation goes to my friends **Dr. Ma'en Al-Jezawi** and **Haitham Raik** for they supported the achievement of this work by their guidance and experience in statistics and programming, respectively.

**Dr. Mohammed Al-Sammarraie** was more than a friend or an office-mate. Without his help, encouragement, and academic discussion, the pace of my work would be slower. I will never forget the nice times we spent in our office.

---

I would like to express my sincere thanks to my special friends **Khaled Magableh**, **Sufian Magableh**, **Rafi Magableh**, and **Aws Magableh**. They have been always there when they were most needed.

Also, I would like to express my appreciation to all my friends, colleagues, and staff in the STRL for the family-like and lovely environment. A special gratitude goes to the technical coordinators **Mrs Lynn Ryan** and **Mrs Lindsey Trent** for their collaboration.



# Contents

<b>Dedication</b>	<b>I</b>
<b>Abstract</b>	<b>II</b>
<b>Declaration</b>	<b>IV</b>
<b>Publication</b>	<b>V</b>
<b>Acknowledgement</b>	<b>VI</b>
<b>List of Abbreviations</b>	<b>XV</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	2
1.2 Motivation and problem statement . . . . .	3
1.3 Scope . . . . .	4
1.3.1 Research scope . . . . .	5
1.4 Research objectives . . . . .	6
1.4.1 Research hypotheses . . . . .	7
1.5 Success criteria . . . . .	8
1.6 Research methodology . . . . .	9
1.7 Thesis structure . . . . .	10
<b>I Background and Literature Review</b>	<b>12</b>
<b>2 Semantic Web and Social Web</b>	<b>13</b>
2.1 Introduction . . . . .	14
2.2 Semantic Web . . . . .	14
2.2.1 Semantic Web ontologies . . . . .	16

2.2.2	Princeton WordNet (PWN) ontology . . . . .	17
2.2.3	MultiWordNet (MWN) ontology . . . . .	21
2.3	Social Web . . . . .	22
2.3.1	Tagging systems and folksonomies . . . . .	22
2.3.2	Tagging for metadata creation . . . . .	24
2.3.3	Folksonomy strengths . . . . .	26
2.3.4	Folksonomy challenges . . . . .	26
2.3.5	Folksonomies and ontologies . . . . .	28
2.4	Summary . . . . .	31
<b>3</b>	<b>Tagging Systems Approaches and Applications</b>	<b>32</b>
3.1	Introduction . . . . .	33
3.2	Statistical and pattern analysis studies . . . . .	33
3.3	Taxonomy of approaches for addressing folksonomies challenges . . . . .	36
3.3.1	Ontological approach . . . . .	36
3.3.2	Social networks approach . . . . .	43
3.3.3	Visualisation Approach . . . . .	47
3.4	Summary . . . . .	49
<b>II</b>	<b>Tagging System Architecture</b>	<b>51</b>
<b>4</b>	<b>Generic Architecture for Tagging Systems</b>	<b>52</b>
4.1	Introduction . . . . .	53
4.2	Standards and criteria for an efficient approach in developing a tagging system . . . . .	53
4.2.1	Integrity of user tags . . . . .	53
4.2.2	Integrity of social interaction patterns . . . . .	54
4.2.3	Rich functionality . . . . .	56
4.2.4	Universality . . . . .	56
4.2.5	Dynamism . . . . .	57
4.2.6	Multilinguality . . . . .	57
4.3	Generic architecture for tag-based systems . . . . .	58
4.3.1	Tagging component . . . . .	64
4.3.2	Searching component . . . . .	69
4.3.3	Semantic component . . . . .	76
4.3.4	Clustering component . . . . .	84

4.3.5	Database component . . . . .	91
4.4	Summary of our architecture . . . . .	93
4.4.1	Virtues . . . . .	93
4.4.2	Limitations . . . . .	93
4.5	Summary . . . . .	94
<b>III</b>	<b>Implementation and Evaluation</b>	<b>96</b>
<b>5</b>	<b>Prototype Implementation of the Tagging System Architecture</b>	<b>97</b>
5.1	Introduction . . . . .	98
5.2	The prototype scope . . . . .	98
5.2.1	Semantic resources . . . . .	99
5.2.2	Social resources . . . . .	107
5.3	Our algorithm for adding system tags . . . . .	109
5.3.1	YouTube . . . . .	109
5.3.2	Database design . . . . .	110
5.3.3	Algorithm implementation . . . . .	112
5.4	Summary . . . . .	122
<b>6</b>	<b>Experiment: Rationale and Design</b>	<b>124</b>
6.1	Introduction . . . . .	125
6.2	The experiment rationale . . . . .	125
6.3	The experiment design and interface . . . . .	126
6.3.1	Introductory page . . . . .	127
6.3.2	Search page . . . . .	127
6.3.3	Results page . . . . .	130
6.3.4	Saving page . . . . .	142
6.4	The database design . . . . .	143
6.5	Sampling design . . . . .	146
6.5.1	Piloting . . . . .	148
6.6	Summary . . . . .	150
<b>7</b>	<b>Results and analysis</b>	<b>151</b>
7.1	Introduction . . . . .	152
7.2	Preparation of the data for analysis . . . . .	152
7.2.1	Data collapsing . . . . .	154

7.2.2	Data Dissection . . . . .	155
7.3	Descriptive statistics . . . . .	161
7.3.1	Collapsed data . . . . .	163
7.3.2	Dissected data . . . . .	163
7.4	Inferential statistics . . . . .	167
7.4.1	Wilcoxon Signed-Rank test . . . . .	167
7.5	The findings and our research hypotheses . . . . .	172
7.6	Summary . . . . .	173
<b>8</b>	<b>Conclusion and Future Work</b>	<b>175</b>
8.1	Research summary . . . . .	176
8.2	Success criteria revisited . . . . .	178
8.3	Contribution to knowledge . . . . .	178
8.4	Comparison with existing related work . . . . .	179
8.5	Limitations and Future work . . . . .	181
<b>A</b>	<b>The Experiment Sample Data</b>	<b>200</b>
A.1	The keywords used to import the YouTube videos . . . . .	200
A.2	Sample data statistics . . . . .	204
<b>B</b>	<b>The Collected Data</b>	<b>209</b>
B.1	The whole data set statistics . . . . .	209
B.2	The collapsed data set statistics . . . . .	212

# List of Figures

2.1	Relations in WordNet [1]. . . . .	18
2.2	Bipolar adjective structure [2]. . . . .	20
2.3	Tag cloud. . . . .	24
3.1	Power law distribution graph [3]. . . . .	35
3.2	FolksAnnotate architecture [4]. . . . .	38
3.3	Merging Social Web with Semantic Web [5]. . . . .	39
3.4	A screenshot from the Del.icio.us page for tag “Pasta” - the inner sidebar shows an expandable hierarchy of related tags. [6]. . . . .	40
3.5	OntoSonomy prototype interface [7]. . . . .	41
3.6	Part of the cluster for the tag “design” [8]. . . . .	44
3.7	A model of tagging system [9]. . . . .	46
3.8	Improved tag cloud [10]. . . . .	47
3.9	A Del.iciou.us visualisation of A tag cloud for Boing Boing website generated by <a href="http://cloudalicio.us">http://cloudalicio.us</a> for the interval (26-10-2009 to 01-11-2009). . . . .	48
4.1	Generic architecture for tag-based systems. . . . .	59
4.2	Tagging system use case diagram. . . . .	60
4.3	Examples of hypernymy/hyponymy semantic relations . . . . .	64
4.4	Generic architecture for tag-based systems - <i>Tagging component</i> . . . . .	65
4.5	Tagging activity diagram. . . . .	66
4.6	Generic architecture for tag-based systems - <i>Searching component</i> . . . . .	70
4.7	Searching activity diagram (using the <i>normalised metadata</i> only). . . . .	71
4.8	Searching activity diagram (using both <i>raw</i> and <i>normalised metadata</i> ). . . . .	73
4.9	Generic architecture for tag-based systems - <i>Semantic component</i> . . . . .	76
4.10	The activity diagram of adding <i>system tags</i> from the semantic ontologies. . . . .	78
4.11	Alternative scenario activity diagram. . . . .	82
4.12	Generic architecture for tag-based systems - <i>Clustering component</i> . . . . .	85

4.13	Clustering activity diagram. . . . .	87
4.14	The activity diagram of adding <i>system tags</i> from the tag clusters. . . . .	88
4.15	The activity diagram of adding <i>system tags</i> . . . . .	89
4.16	Generic architecture for tag-based systems - <i>Database component</i> . . . . .	91
4.17	Logical diagram for the <i>tags</i> tables. . . . .	92
5.1	The components of the tagging systems architecture. . . . .	99
5.2	The semantic component in our prototype. . . . .	100
5.3	The scope of our prototype. . . . .	108
5.4	Diagrammatic representation of the <i>video</i> entity and its attributes. . . . .	110
5.5	Logical diagram for the <i>video</i> entity tables. . . . .	111
5.6	Logical diagram for our videos database. . . . .	112
5.7	Methods diagram. . . . .	114
6.1	Our online environment - Introduction page. . . . .	128
6.2	Our online environment - Search page. . . . .	129
6.3	Our online environment - Results page. . . . .	131
6.4	The layout of the results page. . . . .	133
6.5	Our online environment - Saving page. . . . .	143
6.6	Logical diagram for our entire database (videos data + experiment data). .	144

# List of Tables

2.1	Comparison between traditional and folksonomical classification approaches.	25
3.1	Example of <i>Top-4</i> related tags for some tags [8]. . . . .	44
4.1	Short forms vs. complete form of some English words/phrases. . . . .	62
4.2	Possible probabilities and the retrieved results for the <i>first</i> and the <i>second</i> scenarios. . . . .	74
4.3	Comparison among the three scenarios. . . . .	75
4.4	The addressed challenges in our architecture. . . . .	94
6.1	Some examples of similar <i>user tags</i> and <i>system tags</i> . . . . .	139
7.1	Explanation of the system tags sources and the language of their related user tags. . . . .	155
7.2	The descriptive measures for the whole data set. . . . .	162
7.3	The frequencies of participants' evaluation for both groups. . . . .	162
7.4	The descriptive measures for the collapsed data set. . . . .	163
7.5	The descriptive measures for the <i>source-based dissected</i> data set. . . . .	164
7.6	The descriptive measures for the <i>language-based dissected</i> data set. . . . .	165
7.7	The descriptive measures for the <i>relation-based dissected</i> data set. . . . .	166
7.8	The <i>P-Value(s)</i> of comparing each <i>subset</i> in the <i>source-based</i> dissected data with the participants' evaluation for videos retrieved using user tags.	170
7.9	The <i>P-Value(s)</i> of comparing each <i>subset</i> in the <i>language-based</i> dissected data with the participants' evaluation for videos retrieved using user tags.	171
7.10	The <i>P-Value(s)</i> of comparing each <i>subset</i> in the <i>relation-based</i> dissected data with the participants' evaluation for videos retrieved using user tags.	171

# List of Abbreviation

<b>API</b>	Application Programming Interface
<b>DBMS</b>	Database Management System
<b>FOAF</b>	Friend Of A Friend
<b>HTML</b>	HyperText Markup Language
<b>JAWS</b>	Java API for WordNet Searching
<b>JSP</b>	JavaServer Pages
<b>MWN</b>	MultiWordNet
<b>OOP</b>	Object-Oriented Programming
<b>PWN</b>	Princeton WordNet
<b>SMS</b>	Short Text Message
<b>SPSS</b>	Statistical Package for Social Sciences
<b>SQL</b>	Standard Query Language
<b>UGC</b>	User Generated Content
<b>URL</b>	Uniform Resource Locator



# Chapter 1

## Introduction

### *Objectives:*

---

- Providing an overview of the research problems and motivations.
  - Identifying the scope of the thesis.
  - Presenting the research objectives, questions, and hypotheses.
  - Describing the research methodology.
  - Introducing the thesis structure.
-

### 1.1 Background

The Internet's debut changed the patterns of the daily life for individuals and organisations. The availability of information at a relative ease is the main reason behind the success of the Internet. Nevertheless, the availability of the information has no value unless the information is accessible and retrievable. Therefore, beside the information authoring, information providers were engaged in classifying the information in a suitable way to guarantee that the information is accessible and, thus, survivable.

The process of information classification, or categorisation, is as important as the information generation itself. This process needs a lot of time, money, and effort. In addition, it needs trained people as it has unstable standards. Therefore, it is impractical, to some extent, with the huge amount of information on the Web and the vast number of users who are willing to *consume* the Web content.

At a certain point of the Web evolution, new websites were launched that opened the doors for users not only to *consume* content, but also to *produce* it; so-called User Generated Content (UGC). The real consideration of those websites is who will categorise this massive data, and whether the crowds can categorise the generated content in a consistent way or not.

Tagging was the easiest and most popular solution for data indexing. The principle in tagging is simple and needs no trained users; users add free text words that are best describing their content. Each tag will be considered as a category under which the content is listed.

Nevertheless, tagging is a sword with two edges; it is easy, simple and welcomed by

users, however it produces inconsistent and ambiguous classification of data. Moreover, it suffers from the lack of semantics among tags.

Among the solutions for lack of semantics in tagging systems was the use of the Semantic Web. Hence the name; it is a Web in which data has meanings. Without being expert in Web technologies, it sounds reasonable to use a *web of meanings* to address the problem of a *Web that lacks meanings*.

## 1.2 Motivation and problem statement

With the current success and popularity of the tagging systems (e.g. YouTube), the problem of information browsing and retrieval in such systems becomes a serious challenge. The weakness of information retrieval in tagging systems originates from the inconsistency and ambiguity of tag-based classification of contents. The inconsistency and ambiguity are due to lexical reasons; such as synonymy, polysemy, misspelling, multilinguality, shorthand writing, and others. The gap between the submitted search keywords and the tags used to annotate the contents causes irrelevant results to be retrieved, and most importantly, relevant results not to be retrieved.

A user who is searching in the tagging system for an “*automobile*” cannot find the video, in YouTube for instance, which was tagged by another user using the word “*car*” (synonymous words). Likewise, a user who is searching in the tagging system for a “*baby*” cannot find the photo, in Flickr for example, which was tagged by an Italian user using the Italian equivalent word “*bambino*” (multilinguality). Furthermore, the same problem occurs for the word “*love*” and its shorthand writing “*luv*”, and so forth.

The aforesaid examples illustrate the state of the art in tagging systems. Moving to

another scene in the world of the Semantic Web (so-called Web 3.0), we find ourselves in front of an ideal structure of lexical dictionaries, so-called lexical ontologies. WordNet, for example, is a lexical ontology that presented a “*net of words*” linked to each other based on the semantic relation between the correlated words (e.g. synonyms). MultiWordNet is a similar ontology to aggregate more than one language in one place to support multilinguality.

By moving between the two scenes; the tagging systems and their problems, and the Semantic Web and its lexical ontologies, it seems that the Semantic Web has the solution for the drawbacks of the tagging systems. Indeed, this was the real motive behind the whole work.

Yet, the word “*luv*”, and other shorthand written words, cannot be found in the lexicon. Therefore, it needs to be treated in a different way. If the *context* of such words is defined, the meanings can be extracted (to some extent). Since these words emerged in the tagging systems, then the tagging systems themselves can provide a *context* to define their meanings.

### 1.3 Scope

Tagging is one of the applications that belong to the second generation of the Web evolution, so-called Web 2.0. The main feature of this generation is the empowered role of the user. Web 2.0 applications enabled users to generate and categorise the content, publish their own blogs, socialise via online communities, and build their virtual reality. Consequently, the name “*Web 2.0*” and the name “*Social Web*” are used interchangeably.

Therefore, in this work, we explore the Social Web in general and the tagging sys-

tems in particular. We present an overview about the main concepts, features, advantages, drawback, and the main approaches of research done in this area of knowledge.

Our proposed solution to address some social tagging weaknesses is to exploit the power of the Semantic Web. Hence, an overview about the Semantic Web and ontologies is provided with an emphasis on the lexical ontologies; namely, the WordNet ontology and the MultiWordNet ontology. The potential collaboration between the Social Web and the Semantic Web is discussed as well.

### 1.3.1 Research scope

The metadata used for searching the tag-based systems (e.g. YouTube) is not restricted to tags only. Rather, other kinds of metadata are being used such as title, description, username, etc. In our case, we needed to anatomise the metadata since we are investigating only one kind of metadata; which is the tags. Therefore, tags are considered the only searchable metadata kind in our work.

Even though the tags are considered to be the only searchable metadata, the tags that we are investigating are of two types; the *original user tags* added by real users, and the *new system tags* added by the system. Therefore, another distinction between these tags types is considered.

This research is restricted to address the following tagging challenges only:

- Semantic relations
- Multilinguality
- Shorthand tags

Moreover, we investigate the relatedness between the results retrieved using system tags and the submitted keywords. The time-wise and the space-wise issues are beyond our research scope.

## 1.4 Research objectives

Our proposal for improving the information retrieval in tag-based systems is to add a new set of tags each time the user provides a tag. The new tags will be added by the system in order to overcome the lack of semantics in user tags. Since the new tags will be added by the system, we termed them as “*system tags*”.

The resources from where the system tags are extracted vary depending on the provided user tags. Each added system tag has a relation with the corresponding user tag. If the user tag is a word that exists in the lexicon, its related system tags will be added from the semantic ontologies WordNet and MultiWordNet. Otherwise, the system tags will be extracted from a tag cluster where all tags in each cluster are semantically related.

The purpose of adding system tags is to define meanings (semantics) of the tags in the tagging system. Consequently, the problems presented in Section 1.2 can be addressed. Namely, if there are related contents that are not retrieved because they are not well-annotated, they will be retrieved using the system tags.

### Research questions

The main research questions to be investigated in this research are as follows:

*Question 1:* Are there related results that will be retrieved from tag-based systems by searching using the new **system tags** only?

*Question 2:* What is the difference, in terms of relatedness, between the results retrieved by using the **user tags only** and the results retrieved by using the **system tags only**?

### 1.4.1 Research hypotheses

In order to guide our research process and to identify the right kind of data that we need for our investigation, we make some hypotheses that should be tested by further investigation. “A *hypothesis* is a logical supposition, a reasonable guess, or an educated conjecture that provides a tentative explanation for a phenomenon under investigation” [11]. Indeed, the hypothesis itself is not normally tested to be supported or rejected. Rather, its logical opposite or negation, so-called the *null hypothesis*, is tested [12]. To support the hypothesis, we strive to reject the null hypothesis. That is; the original hypothesis is accepted if the null hypothesis has been rejected. The original hypothesis, that we are primarily interested in, is now called the *alternative hypothesis*.  $H_1$  is the symbol used to represent the alternative hypothesis, whereas  $H_0$  is the symbol used to represent the null hypothesis [12].

Based on the research questions abovementioned, we could formulate our research hypotheses and, obviously, the null hypotheses. Each research question has a corresponding hypothesis. The following is the first hypothesis, and its null hypothesis, for the first research question (Question 1):

The first hypothesis:

H1: Adding system tags as metadata can retrieve results that are related to the searching keywords when searching in tag-based systems

The first null hypothesis:

H1<sub>0</sub>: Adding system tags as metadata can **NOT** retrieve results that are related to the searching keywords when searching in tag-based systems

For the second research question (Question 2), here are the alternative hypothesis and its null hypothesis:

The second hypothesis:

H2: The degree of relatedness between the results retrieved using system tags and the search keywords is **the same as** or **higher than** the degree of relatedness between the results retrieved using user tags and the search keywords

The second null hypothesis:

H2<sub>0</sub>: The degree of relatedness between the results retrieved using system tags and the search keywords is **lower than** the degree of relatedness between the results retrieved using user tags and the search keywords

## 1.5 Success criteria

Supporting or rejecting the abovementioned hypotheses verifies whether the system tags can improve the information retrieval in tagging systems or not. As aforesaid, the hypotheses investigate the relatedness between results retrieved using system tags and search keywords. As relatedness is a subjective criterion, it needs to be compared on two different sets of results where one set is retrieved using system tags and the other set is retrieved using another kind of metadata, for the same subjects under the same conditions. Therefore, we decided to compare the relatedness of results retrieved by using the system tags we proposed with the relatedness of results retrieved by using tags that were provided by



real users on YouTube. Other conditions of comparison were fixed in the two compared cases.

The system tags will be considered a successful solution if the results of the comparison reveal one of the following cases:

- **The relatedness in both cases is the same:** This case indicates that the new system tags are as valid as the user tags with more coverage of semantically related results.
- **The relatedness for system tags case is higher:** This case indicates that the new system tags are more valid than the user tags with more coverage of semantically related results.

## 1.6 Research methodology

The following work packages summarise the methodology followed in this research:

- **Research background:** The research started by reviewing the literature in the area of Social Web and Semantic Web. After acquiring the required background, we came up with a novel approach for integrating the Social Web and Semantic Web technologies to address some of the existing shortcomings in tagging systems. Furthermore, we identified a set of criteria for an efficient approach in developing a tagging system.
- **Generic architecture:** Having reviewed the literature in the scope of this research, a generic architecture was developed which complies with the criteria we developed in the *research background* work package. The architecture gives directions for addressing various tagging challenges. It consists of five components; tagging component, searching component, semantic component, clustering component, and database component.

- **Prototype implementation:** Our prototype implements the semantic component, clustering component, and the database component of the generic architecture. Real data set was imported from YouTube tagging system and stored in our database. The relevant system tags were added from semantic resources (WordNet and MultiWordNet) and social resources (Flickr clusters). Afterwards, an online interface was designed to conduct an online experiment where real users can browse and search our database (data set).
- **Evaluation:** A big sample (204 subjects) of users were asked to search in the online environment we designed, and to evaluate the relatedness between the retrieved videos and the submitted keywords using a Likert scale. 1,391 videos were retrieved and evaluated by using user tags and system tags. A statistical approach was followed in order to compare the participants' evaluation of the videos retrieved using user tags with the participants' evaluation of the videos retrieved using system tags. A well-known inferential statistical test called *Wilcoxon Signed-Rank* was used to detect whether there is a significant difference between the evaluations of the results in the two groups.

## 1.7 Thesis structure

Including this chapter, this thesis contains eight chapters. The following is a summarised description of these chapters starting from the next chapter.

- **Chapter 2:** This chapter provides an overview of the Semantic Web and its ontologies, and the Social Web and its tagging systems. The chapter explores how these two Web generations can collaborate.
- **Chapter 3:** This chapter presents the related work in the area of tagging systems. The proposed solutions for tagging challenges are classified in this chapter under

three main categories; ontological approach, social networks approach, and visualisation approach.

- **Chapter 4:** This chapter introduces the aforementioned set of criteria we identified. Moreover, it illustrates our generic architecture for tagging systems and discusses the functionalities of its five components.
- **Chapter 5:** The prototype implementation of the main components of our generic architecture is discussed in this chapter. It describes the process in which real data (with user tags) is imported from the YouTube. The chapter also presents our algorithm for adding system tags from semantic and social resources. Furthermore, this chapter discusses the rationale for deciding which kind of system tags can be added.
- **Chapter 6:** In order to evaluate the prototype system we implemented, we needed to develop an online environment to enable a sample of users to retrieve and evaluate information from our prototype system. This chapter describes the online environment design and interface. Moreover, it presents the design of the database where the collected evaluations were stored. A description of the sampling design and the pilot study we carried out is provided.
- **Chapter 7:** Having collected the participants' evaluation, we prepared the collected data for statistical analysis. Two types of statistics are provided in this chapter; *descriptive statistics* and *inferential statistics*.
- **Chapter 8:** Summaries, conclusions, and potential future work of our research are mentioned in this chapter.

## **Part I**

### **Background and Literature Review**

# Chapter 2

## Semantic Web and Social Web

### *Objectives:*

---

- Giving a background about the Semantic Web and its ontologies. Specifically, Princeton WordNet ontology and MultiWordNet ontology.
  - Giving a background about the Social Web, tagging, and folksonomies.
  - Discussing the properties of metadata created in tagging systems.
  - Presenting the advantages and the disadvantages of tagging systems.
  - Discussing the trade-off between Web 2.0 and Web 3.0.
-

### 2.1 Introduction

As our work comprises two generations of the Web, this chapter provides an overview of the Semantic Web and its ontologies, as well as the Social Web and its folksonomies. In the light of the success of folksonomies in engaging end-users to generate data as well as metadata, we shed some light on the folksonomies challenges to identify them and to be aware of their nature and origin.

In order to support the claim that ontologies and folksonomies can coexist, the benefits and trade-offs between them is discussed at the end of the chapter.

### 2.2 Semantic Web

The Internet is becoming an essential pillar of our modern world. It almost penetrates every side of our daily life; it is used for many purposes including academic research, entertainment, communication, commerce, banking, and others. The use of Internet, as well as the information on the Internet, is increasing astronomically. This information is presented mainly via natural languages, and it is not labelled in a meaningful manner for computers to understand. Computers do not understand the natural languages although they read these languages. Moreover, the incapability of computers to access, process, and interchange this information in understandable manner is reflected on the users [13]. Users face problems in searching information and resources discovery; many irrelevant results are retrieved, and more importantly, many relevant results are not retrieved [13]. This created the motivation to think about a Semantic Web.

There is a need for a Web infrastructure that can integrate and synchronise information on the Web. For instance, an update on one website will be immediately reflected

on other related websites. One of the most common ways to present this integration is to combine the information of both websites in one relational database. This is not applicable because it is unlikely for independent organisations to have a single database. The information on the Internet belongs to many parties (millions), so it is impossible to put it in one relational database [14]. Nevertheless, we yearn to integrate all data resources on the Web in a machine understandable manner, so that all the data in the world look like one huge distributed database. In fact, this is the so-called Semantic Web vision [15].

Tim Berners-Lee, the inventor of the World Wide Web, was the first who coined the term “*Semantic Web*” at the end of the last century. His definition of the Semantic Web is “*a Web of data that can be processed directly or indirectly by machines*” [15].

The Semantic Web is the current Web, plus meanings of data; it is therefore an extension of the current one [16]. The Semantic Web applications do not focus on the presentation but on the subjects of presentation. In other words, semantic applications will explicitly define the subjects, and determine the underlying relationships between these subjects; therefore, they can generate the presentation as needed [14]. Semantic Web is not about links between Web pages; Rather, it describes the relationships between things (like A is a part of B and Y is a member of Z) and the properties of things (like size, weight, age, and price) [17].

Currently, many parts of the Semantic Web are already existing, as well as many semantic ontologies and Semantic Web languages that power the vision and facilitate the development of the new machine understandable Web [16].

### 2.2.1 Semantic Web ontologies

Ontologies represent the key enabling technology for the Semantic Web vision. Ontology has been originated as a branch of philosophy which concerns with articulating the nature and the structure of the world. The concept, later on, was borrowed from philosophy and is used in information technology. Ontologies were developed first in artificial intelligence to facilitate knowledge sharing and reuse. Then, ontologies have become one of the active research topics in many fields such as knowledge management, knowledge representation, knowledge engineering, natural language processing, intelligent information integration, cooperative information systems, information retrieval, and electronic commerce [18]. This popularity for ontologies is due to their promise of *a shared and common understanding of a domain that can be communicated between people and applications* [13, 18]. This promise is harmonised with the Semantic Web vision: *common understanding that can be understood by both human users and software agents*.

#### Ontology definition

Many definitions were proposed to define what an ontology is. Within the context of information technology sciences, Gruber has defined ontology in 1993 as “*an explicit specification of a conceptualisation*” [19]. Conceptualisation is a simplified explanation of domain concepts and their relations. Normally, we have conceptualisation of things in our minds; which is *implicit* conceptualisation. In ontologies, this conceptualisation should be specified *explicitly*. Later in 1997, this definition was expanded to add a new dimensions, Borst defined the ontology as “*a **formal** specification of a **shared** conceptualisation*” [20]. Thus, Borst emphasises the notion of agreement on the conceptualisation which will facilitate the reuse of ontology; “*shared*” means a consensus among several parties. “*Formal*” means that it has a precise notation. Studer merged the two definition in 1998 in one definition: “*An ontology is a formal, explicit specification of a shared concep-*



tualisation” [21]. Therefore, the latter definition states that the shared conceptualisation must be formal and explicit.

One of the well-known existing ontologies that is commonly used in different applications is the *Princeton WordNet Ontology*. The WordNet ontology is essential part of our work.

### 2.2.2 Princeton WordNet (PWN) ontology

Most of the online dictionaries today were produced to be understood by humans, not by machines. Further, WordNet is a lexical ontology, created by a team of researchers at Princeton University, that can be understood by both humans and machines. It offers a combination of traditional lexicographical resource and modern semantic ontology. The WordNet dictionary<sup>1</sup> contains English nouns, verbs, adjectives, and adverbs all organised into sets of synonyms, so-called *synsets*. The synsets are linked to each other by semantic and lexical relations. The semantic relations are between word meanings, while the lexical relations are between word forms. Two words, or more, that share the same meaning are said to be *synonyms*. A word is *polysemous* if it appears in different synsets with different meanings, each meaning represents a possible *sense* of the word. Circa 17% of the words in WordNet are polysemous, while around 40% have one or more synonyms [1, 22, 23].

WordNet is rich of information about semantic and lexical relations between words. It is a trial to model the lexical knowledge of the native English speakers [24]. Table 2.1 shows the main relations in PWN. These relations are:

- *Synonymy relation* [1, 22, 24, 25]: The most important relation in the WordNet is the similarity of meaning. According to [22], two words are considered to be

---

<sup>1</sup>We will use the words ontology, dictionary, or database to describe the WordNet throughout this work.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponymy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs		

Figure 2.1: Relations in WordNet [1].

synonyms if the substitution of one for the other in a linguistic context will seldom alter the meaning in that context. In PWN, synonymy is implicitly presented in the inclusion of the words with the same part-of-speech in the same synset.

- *Antonymy relation* [1, 22]: According to [22], “The antonym of a word  $x$  is sometimes *not- $x$* , but not always”. That is; *rich* and *poor* are antonyms, but it is not necessary that the *not-rich* is *poor*, nor vice versa, the *not-poor* is *rich*. It is a bit complex to define the antonymy relation, but generally speaking it is the *opposing name*. This relation has special importance in organising the meanings of adjectives and adverbs.
- *Hyponymy relation* [1, 22, 26, 27]: Hyponymy is a transitive relation that refers to the sub-name (is-kind-of) of a given noun. For example *tree* is a hyponym of (is-kind-of) *plant*. Thus, it is a specialisation relation between a specific and more general word. The inverse of this relation (the generalisation) is called “*hypernymy*”. Therefore, *plant* is the hypernym of *tree*. This relation between word meanings organises the nouns into a hierarchical structure. Therefore, a given word inherits the

super-ordinate's properties.

- *Meronymy relation* [1, 22]: Meronymy is a complex relation that refers to the part-name of a given noun. *Limp* is a meronym of (is-part-of) *tree*. In WordNet, there are three types of this relation; component parts, substantive parts, and member parts.
- *Entailment relation* [26, 28, 29]: Entailment relation is a unilateral relation, that is, doing one *verb* entails doing the other one, but not the other way around; *snoring* entails *sleeping*, but *sleeping* does not entail *snoring*. There are more than one type of entailment between English verbs <sup>2</sup>.
- *Troponymy relation* [1, 26, 28]: Troponymy relation is a special kind of entailment relation between verbs. Similarly to the hyponymy relation between nouns, troponymy relation is between verbs although the hierarchy in verbs is shallower. The *is-kind-of* relation between nouns is comparable to the *is-manner-of* relation between verbs. For example, *walk* is a troponym of (is-manner-of) *move*. Necessarily, *walk* entails *move* also.

There are some other relations that revolve around the abovementioned six main relations. Some of these salient relations are:

- “*Similar to*” relation [2, 30, 31, 32]: WordNet contains two types of adjectives; descriptive and relational. Descriptive adjectives ascribe a value (e.g. *big*, *heavy*) of an attribute (e.g. *size*, *weight*) to a noun. Relational adjectives are related to, pertained to, or associated with some noun (such as *presidential*, *managerial*, *nuclear*).

Some descriptive adjectives are similar in meanings but not close enough to put together in one synset (e.g. *moist* and *wet*). Furthermore, the antonymy relation is very important in the classification of adjectives in Wordnet. Some descriptive

---

<sup>2</sup>For more information, see [28].

adjectives are antonymous (e.g. *heavy/light* and *weighty/weightless*), while some others are not (e.g. *ponderous, massive, airy*). In these cases, the solution in WordNet was to link these adjectives together using *similar to* relation. Organising the non-antonymous adjectives in clusters around antonymous adjectives gives the former an indirect antonym via the latter. As seen in Figure 2.2, the head adjective is antonymous, while the satellite one is non-antonymous. The descriptive adjective *moist* does not have *direct* antonym. But it is similar to the adjective *wet*, which has direct antonym *dry*. So, the *indirect* antonym for *moist* is *dry*, and so forth.

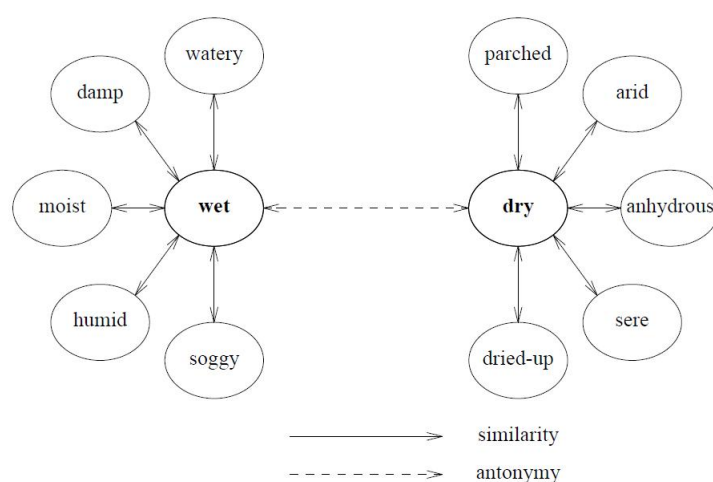


Figure 2.2: Bipolar adjective structure [2].

- “*Also-see*” relation: This relation exists in WordNet, the WordNet Application Programming Interfaces (APIs), and is used in many applications that use the WordNet. Mutually, it is called *related term*. But unexpectedly, we could not find in the literature the criteria used by WordNet to decide whether two synsets are linked via this relation. According to [31, 33], possibly human judgment was used to make this decision on a case-by-case basis. For example, *hostile* is linked to (*aggressive, hateful, offensive, unfriendly, unpeaceful, violent*) with *also-see* relation.

Due to its accessibility and quality, WordNet has become the ideal tool for many applications such as semantic tagging, information retrieval and much more [34].

### 2.2.3 MultiWordNet (MWN) ontology

MultiWordNet is a multilingual database designed based on PWN structure. This project has been created at ITC-irst aiming at building an Italian WordNet that is strictly aligned with PWN. MWN has been built following a model called *Expand Model* which entails building the language-specific WordNet and importing the maximum possible semantic relations from PWN. That is; if there is a relation holding between two synsets in PWN, the same relation will hold between the corresponding synsets in the Italian WordNet whenever possible [35].

Some of PWN relations are common to all languages whilst others are language-dependent. *Semantic relations* are common ones (e.g. hypernymy, entailment, etc), whereas the *lexical relations* are language-dependent. The current MWN contains only two languages, thus, the information in MWN is contained in three main modules; *common-database* module, *English-database* module, and *Italian-database* module. The *common-database* module contains the *semantic relations* between synsets which hold for all languages. The *Italian-database* and *English-database* modules (language-specific database) are similar in their structure but different in the data they store; each language-specific database contains the information and *lexical relations* for a specific-language [35].

The cross-language linkage is realised by using the same identifier for the corresponding synsets in the different languages. For example, “*gatto*” is the Italian translation for the English word “*cat*”, therefore, they are stored in two different tables but share the same identifier (n#01630731).

## 2.3 Social Web

Online social networks are one of the main elements of the Social Web, so-called Web 2.0. The term Web 2.0 was coined by Tim O'Reilly referring to the new Web generation with new usage patterns in the online world. This generation created a platform for intense communication, social interaction, and user-generated websites where like-minded people meet and collaborate. This new trend in the Web management transferred the Web from *read-only* Web to *read-write* Web. In the read-write Web, end-users are producing the Web content rather than just consuming it [36].

The first wave of the Social Web was due to the appearance of Web-based communication and collaboration forms such as blogs, wikis, and other online social networks. This phenomenon allowed the users to have their own space on the Web at relative ease [36]. Nowadays, millions of users are storing and sharing their knowledge online in a searchable style. The Social Web provides a sustainable fountain of publicly available electronic content that reflects the wisdom of crowds, and therefore the Social Web presents a collected knowledge system [37].

The social networks have a variety of purposes; for example, some of them represent a place where friends and families can meet, communicate, and share their events (e.g. Facebook), other networks are to aggregate people who have a common interest about specific field of knowledge (e.g. Bibsonomy), and others are for knowledge dissemination and sharing (e.g. Wikipedia), etc.

### 2.3.1 Tagging systems and folksonomies

Tagging is the process where a user creates and manages metadata in a form of free-text keywords (so-called tags) for community-shared resources in order to share, describe, an-

notate, index, and categorise these resources via a Web-based interface [7, 38, 39, 40, 41, 42, 43, 44, 45, 46]. If these resources are permitted to be tagged by more than one user, then it is a *collaborative tagging*. The name “*collaborative tagging*” has many alternative names that are used interchangeably. One of the most common used names is “*folksonomy*” that was coined by Thomas Vander Wal [47]. The word itself is not an English word. Rather, it is a portmanteau that came from two English words; *folks* and *taxonomy* [4, 41, 48]. So, folksonomy is a taxonomy (classification) that is made by folks (people). Collaborative tagging reflects the common understanding of a certain resource from the users’ point of view [43]. More to the point, collaborative tagging is known also as social taxonomy, social classification, social indexing, and social tagging [7, 42, 44].

Each tag, in tagging systems, will represent a category under which the resources will be classified. And hence, the same resource can appear under many categories. The tag is a main element of the tripartite model of: actor (tagger), instance (tagged object), and the tag itself [49].

Tag popularity refers to the level of use frequency of that tag by the users [45]. The most popular tags are often depicted in a “*tag cloud*”. A tag cloud is a Web-based visual representation of the social tags, used to support navigation and retrieval of tagged data. It displays the tags as eye-catching hyperlinks in paragraph-style layout, usually in alphabetical order, with different font attributes such as colour, weight, and size. The font attributes represents the frequency of tags’ use [50, 51, 52, 53]. Figure 2.3 shows an example of a tag cloud.

One of the most well-known applications of collaborative tagging is the social bookmarking. Bookmarking is the practice of saving the favourite website that users wish to visit in the future onto the computer hard drive [54]. The social bookmarking is when

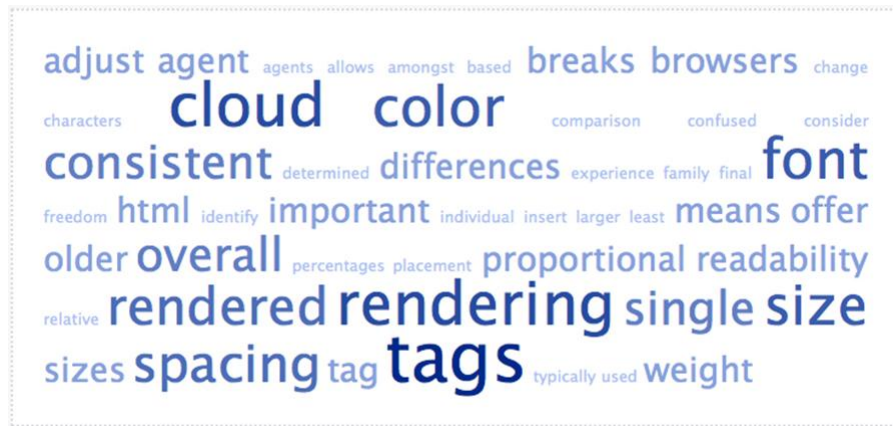


Figure 2.3: Tag cloud.

the users register to a website to save, tag, and share these bookmarks for the purpose of future personal and public use. This phenomenon started few years ago when websites like Del.icio.us appeared.

Collaborative tagging in general, and social bookmarking in specific, becomes an interesting area of research in the fields of information retrieval, data mining, knowledge extraction, ontology learning, and Web intelligence, since it provides a huge amount of user-generated annotations that reflect the interest of millions of users [54, 55, 56].

### 2.3.2 Tagging for metadata creation

Tagging systems enabled users to participate in metadata creation. Here we group the metadata creation approaches and methods into two main types; traditional metadata creation approach and folksonomical metadata creation approach.

- **Traditional Metadata Creation Approach:** In this approach, metadata is created by professionals or authors in the form of catalogue records [57]. The problems of this traditional approach are the continual evolution of new standards and the need for trained staff for the categorisation process [58]. More or less, there is a consensus about the difficulty and high cost of the traditionally metadata creation



in terms of effort, time, and money [45, 58, 59, 60, 61, 62, 63]. Furthermore, the traditional human-created metadata is error-prone, and more likely to be inconsistent. The inconsistency normally happens due to the variations in the cataloguers' judgment over time [60].

- **Folksonomical Metadata Creation Approach:** In tagging systems, the metadata is created by the end-users in the form of free tags. Rather than a hierarchical or exclusive classification, the classification of the contents in folksonomy is relaxed, flat (non-hierarchical), and inclusive [41, 48]. It has no constraints of a predefined taxonomy like the traditional cataloguing [43].

We built Table 2.1 below to compare between the traditional and folksonomical approaches. It explicitly shows the differences, strengths, and weaknesses of both.

Criteria		Traditional Approach	Folksonomical Approach
Cost	Effort	Difficult	Easy
	Time	Time consuming	Very quickly
	Money	Expensive	Almost free
Indexer		Professionals	End-users
Structure		Hierarchical structure (top-down)	Flat structure (bottom-up)
Inclusivity		Exclusive	Inclusive
Standards		Unstable standards	Fixed or no standards
Training and Knowledge Requirement		It needs trained and professional staff	Anyone can do it
Compatibility with the Web		Impractical with huge amount of data on the WWW	Compatible
Scalability and Flexibility		Hardly scalable	Scalable and flexible
Ambiguity and Inconsistency		Exist, but less than the folksonomical approach	Exist, but more than the traditional approach

Table 2.1: Comparison between traditional and folksonomical classification approaches.

### 2.3.3 Folksonomy strengths

The following are the advantages of folksonomies:

- Ease of use; it is very quick, simple and straightforward. Moreover, users can tag without formal training [64].
- Cost effective way of metadata creation; and thus, of information classification and categorisation. This new way is more flexible and scalable as it handles the growing amounts of data in a graceful manner. Furthermore, it is compatible with the vast amount of contents on the Web.
- Reflection of the users' understanding of the data, not the authors' nor the librarians' understandings [43, 56]. Moreover, it reflects the vocabularies used by end-users [57, 65].
- Dynamicity in adaptation to changes that emerge in users' vocabulary [66].
- Good recommendation system for the like-minded people [41, 67, 68, 69, 70].
- Effective content management systems (in particular, the social bookmarking websites) [71].
- Multidimensional classification of content; one tagged resource can belong to different categories depending on the tags [64, 65].
- Provision of information about the users' needs, habits, areas of interest, and how these interests are being described [64].

### 2.3.4 Folksonomy challenges

By analysing the current collaborative tagging systems, it is notable that the main prominent challenges are ambiguity, inconsistency, and redundancy [44, 48, 57, 70, 72, 73, 74,

75, 76, 77]. This is normal since the collaborative tagging systems (by their nature) are shared by many users. These users came from different backgrounds, cultures, countries, domains, and tongues. The diversity and variety of the users' behaviours would inevitably create inconsistent tags that would give ambiguous identification of the tagged objects.

The ambiguity and inconsistency of the tags came mainly from linguistic reasons. The following is a description of these linguistic reasons:

- Word synonyms [7, 40, 41, 44, 46, 48, 57, 67]: There are many objects that have different words (sometimes verbs) to identify them. For example, the word *fair*, and the word *exhibition* can be used to describe the same thing. In this case we had one meaning and different words.
- Word polysemy (homonym) [7, 40, 41, 44, 48, 67, 78]: Here, the case is the opposite of synonymy; we have one word, but different meanings. Back to the example above, the word *fair* is polysemous as it has the meanings of exhibition, blonde, just, reasonable, beautiful, sunny, unblemished, favourable, thorough, legible, etc.
- Lexical forms [7, 40, 41, 48, 79]: The taggers use different lexical forms such as singular words, plural words, conjugated words, active verbs, or passive verbs.
- Alternative spellings [48]: If we take the English language as an example, we note that there are British and American spelling for some words. *centre* is the British spelling, while *center* is the American one for the same word. *Favourite* and *favorite*, and *organisation* and *organization*, are other examples. Such a problem can be considered as an easy or simple one as the alternatives for the word are very limited. And thus, can be easily processed and programmed.
- Misspelling errors [44, 48, 79]: The taggers are humans. This implies the possibility of some mistakes. These mistakes will be understood as new words or new tags

by the computers, while in fact they are not. This problem is simple as it can be addressed by including a spelling-checker in the tagging system.

- Badly encoded tags [79]: In some cases, users group words in an unlikely way (e.g. TimBernersLee). Including the mechanism used in some text editors (e.g. Microsoft Word) can participate in solving such a problem.
- Specialised tags [7, 79]: Some taggers use special terms that are considered as “*nonsense*” tags. This type of tags is normally shared among a group of friends or co-workers. These tags are meaningful and understandable only among the groups’ members, but it has no meaning to the wider community.
- Key phrases instead of keywords [49, 57, 67]: Normally, the tags are separated by spaces in most tagging systems. Nevertheless, few tagging systems allow spaces in tags. If the tagger uses spaces in a folksonomy that does not allow spaces; then it is crucial as the meaning intended by the whole words together as one tag would not be the same when these words are separated into multiple tags.
- Different languages [4, 72, 80, 81]: The Web is distributed and open to everybody, taggers come from different continents and thus they have different tongues. Moreover, due to English language globalisation, some of non-English speakers tag using two languages at least; their native language and English language.

### 2.3.5 Folksonomies and ontologies

In fact, by talking about folksonomies and ontologies, we refer to the latest two generations of the Web evolution respectively; Web 2.0 and Web 3.0, or namely, the Social Web and the Semantic Web. Here we discuss where the Social Web and the Semantic Web meet or, more likely, whether they meet or not, and whether they are alternatives or complements to each other. This debate mainly appeared when Shirky [82] claimed that the

idea of ontology is overrated, and folksonomies present an interesting alternative for the controlled vocabulary of Semantic Web ontologies. Few researchers adopted this belief [41, 83]. On the other hand, many researchers believe that tagging is not an alternative for the controlled vocabulary [4, 5, 7, 37, 39, 42, 48, 65, 79, 84, 85, 86, 87, 88, 89, 90, 91]; each of them can exploit the power of the other one for advantages.

Semantic Web studies the knowledge representation on the Web in such a way that some semantics are associated to the data; therefore it becomes machine understandable. On the other hand, Social Web annotates the Web contents and creates the metadata, which implies the users' participation in knowledge representation, organisation, and categorisation on the Web. Obviously, both the Social Web and the Semantic Web participate in the knowledge management and representation. Indeed, the ideas of Web 2.0 and Web 3.0 are not exclusive alternatives [36]; rather, they are essentially compatible and can co-exist [92]. Furthermore, from a historical perspective, the Semantic Web was originally expected to be filled by users' annotations [36].

### **What Web 2.0 can offer to Web 3.0**

The folksonomies' collective categorisation represents the social knowledge that can be used as an initial knowledge base for constructing ontologies. In particular, the extraction of ontological structure from the folksonomies can reduce the effort needed by human authors, and come up with simpler ontology. Dynamicity is another valuable contribution offered by folksonomies to the Semantic Web; folksonomies are changing over time, requiring mechanisms for authors to capture the changing properties of the modelled domain, and apply them to the corresponding ontologies. This way, the ontologies will be up-to-date and will have a dynamic social trait [92].

The Social Web offers a collective knowledge system. The collective knowledge sys-

tem is a human-computer social system in which machines can collect large amounts of human-generated knowledge, and search engines can retrieve the information stored in that system by querying strategies tuned to the content generation processes. Such a system can provide collected intelligence, but not collective intelligence. The difference is that the latter produces *emergent* knowledge that is not intentionally formed by the human contributors themselves. Web 3.0 can exploit the Web 2.0's collected intelligence by modelling it in a semantic way, so that by applying reasoning methods one can come up with true collective intelligence that facilitates new knowledge creation [37].

By experience, users are willing to provide content as well as metadata on the Social Web via registered users, which offers rich information about the users' profiles. Web 3.0 can exploit this significant information to match users with similar interests which might be helpful for recommendation systems [36].

### **What Web 3.0 can offer to Web 2.0**

One of the weaknesses of the Social Web is that the data is not machine understandable, and thus, not machine processable. The ontologies derived from folksonomies can articulate the collected social knowledge in a machine processable form. This will significantly improve the information retrieval enabling the Web 2.0 search engines to enhance the results quality [92].

Unstructured, ambiguous, and inconsistent tags in the folksonomies can be efficiently utilised once they are structured. The Semantic Web provides a suitable standard infrastructure that can be exploited to structure the aggregation of the social knowledge, and consequently, to boost the data integration and exchange in the Social Web [36].

Current Social websites are isolated from one another. The potential interoperation

among many social communities is expensive due to the lack of compatibility since they use heterogeneous data representations [93]. However, the Semantic Web offers a solution to make social websites interoperable. Indeed, some Semantic Web developments can be applied to the Social Web to enable data portability among the social networks by using so-called Friend-of-a-Friend (FOAF) and Semantically Interlinked Online Communities (SIOC) ontologies [93, 94].

## 2.4 Summary

As the core of our work throughout this thesis is about the use of Semantic Web technologies to improve the information retrieval in the social tagging systems, we presented a background about the semantic and the social technologies.

With the current Web, there are several successful and commonly used social applications that engaged high numbers of users in the Web content generation and classification. Tagging systems allowed users to participate in creating the metadata, beside the data itself, in a low-cost, fast, and easy manner. Although this phenomenon is widely spread and accepted, tagging systems suffer from drawbacks that need to be addressed in order to improve the users' interaction with such systems. The lack of semantics in social tagging systems is a prominent challenge. Therefore, the semantic ontologies WordNet and Multi WordNet, for example, might help in addressing this challenge.

Web 2.0 and Web 3.0 can collaborate to improve the end-users' experience on the Web. They are complementing each other; none of them is an alternative for the other.

## Chapter 3

# Tagging Systems Approaches and Applications

### *Objectives:*

---

- Discussing studies that explore the anatomy of tagging activities.
  - Presenting a taxonomy of related work approaches.
  - Reviewing some related work that tried to address the folksonomy challenges.
-



### 3.1 Introduction

Several researchers have tried to address the challenges of folksonomies. The methods developed by these researchers vary in terms of the used techniques and tools. This chapter reviews these methods and classifies them into three main approaches.

Before reviewing the proposed solutions of the folksonomy challenges, a diagnosis for the tagging behaviours and patterns should be carried out. To this end, this chapter starts by introducing studies that analysed the tagging practices and usage patterns in folksonomies. These studies follow a statistical approach to explore latent phenomena in tagging activities.

### 3.2 Statistical and pattern analysis studies

Such studies give the basis of addressing the ambiguity and inconsistency in folksonomies rather than addressing these problems directly. They investigate the tags frequency, ranking, and popularity. Moreover, they explore the relations between tags, tagged objects, and user profiles [44]. Such studies give a coherent understanding of the tagging activities' properties such as patterns, behaviours, distributions, formation, and stabilisation.

1. *Folksonomy Formation and Stabilisation study* [41, 44]: This study aimed at analysing the structure and the dynamic aspects of folksonomies, with Del.icio.us as a case study. The study found regularity in user activity, tag frequencies, kind of tags used, bursts of popularity in bookmarking, and stability on proportions of tags within a given tagged resource. The study revealed that the tags for any object are classified in different categories according to the information these tags convey and how they are used. Here are these categories:

- Identifying what (or who) it is about; such as *Britain*.

- Identifying what kind of thing a tagged item is; such as *country* and *book*.
- Identifying who owns or created the tagged content; such as *ministry* and *embassy*.
- Identifying subjective characteristics of the tagged objects; such as *funny* and *beautiful*.
- Identifying the content in terms of its relation to the tagger, such tags starts normally with the pronoun *my*; such as *mycar*.
- Identifying the tagged objects according to a task. For example when collecting information related to performing a task, that collected information (result pages) may be tagged according to that task; such as *toread* or *jobsearch*.

The study concluded that most of del.icio.us users were tagging for the purpose of personal use rather than public benefit. Nevertheless, social bookmarking systems, such as del.icio.us, can represent a good recommendation systems.

2. *Collaborative Tagging Patterns and Inconsistencies study* [95]: This study explored the tagging frequency and co-word analysis metrics. Co-words analysis detects the number of times each pair of words (tags) occur together to see the relationship between the words that co-occur frequently. This study was conducted using real data from del.icio.us bookmarking system to measure the similarity between the individual users' classification and the traditional ways of document classification and indexing.

The findings suggest that the first look at the tags in a folksonomy gives a chaotic impression, but after simple exploration of these tags, regular patterns can be observed. These patterns are consistent to some degree with conventional indexing. Furthermore, it occurs frequently that synonymous words are existing in the tag lists for a tagged resource. Hence, a kind of semantics exists in folksonomies. Moreover,

the study showed that 16% of the studied tags were time-related tags; this temporal dimension of the users' classification adds a flexibility and time-sensitivity that did not exist in the traditional classification schemes. On the other hand, time-related tags might retrieve confusing results at a later time.

3. *Collaborative Tagging Dynamics study* [73, 96]: This study uses Del.icio.us data to examine the stability of tag frequencies distribution. Finding how much this distribution is stable indicates the degree of users' consensus about the optimal tags to describe particular resources. As revealed in the studies [6, 71, 76], this study emphasised that the distribution of tags follows the power law distribution; a small number of tags are used in a high frequency, and a high number of tags are used in low frequency (see Figure 3.1).



Figure 3.1: Power law distribution graph [3].

The power law distribution graph has a “*long tail*” and a “*short head*”. The latter suggests that there is a consensus among the users about the tag-based categorisation of information. Therefore, the tags that are in the “*short head*” have semantic relations among each other.

From the statistical and pattern analysis studies, we argue that folksonomies contain, to some degree, valuable semantic relations. Several studies emphasised the power law

distribution of tags for a given resource. The power law distribution indicates a consensus among folks on the meanings of tags. Therefore, one folksonomy can benefit from the semantics of another folksonomy if they collaborate, and this can be augmented by increasing the number of collaborating folksonomies.

### **3.3 Taxonomy of approaches for addressing folksonomies challenges**

By reviewing the literature of tagging systems, there are many attempts to address their challenges. The researchers followed different approaches to leverage the data classification, and thus, the information retrieval in folksonomies. One approach was to capture the power of the Social Web; either by using the data contained in the folksonomy or by aggregating more than one folksonomy together. Another approach was to capture the power of the Semantic Web; either by using domain ontologies or lexical ontologies. A third approach was to work on intuitive visualisation of the Graphical User Interface (GUI) of the folksonomy, so that the browsing experience becomes more meaningful. Although there are no rigid edges among these approaches, we made every effort to classify the attempts into three main approaches<sup>1</sup>:

- *Ontological approach*
- *Social networks approach*
- *Visualisation approach*

#### **3.3.1 Ontological approach**

In the context of ontological approach, we differentiate between two types of ontological methods that have been investigated in the area of tagging systems; *building an ontology*

---

<sup>1</sup>This classification was guided by [44].

*for folksonomy, and using other domains' ontologies in the folksonomy.*

### **Building an ontology for the folksonomy**

This category of methods conceptualises the folksonomies by building an ontology for the folksonomy. It defines the main entities in the tagging systems, the relations among these entities, and their properties. Actually, these methods have the basic tripartite: *tag*, *tager*, and *tagged item* with slight differences. The resulting ontology can help in data exchange among different folksonomies that use heterogeneous tagging data representations. We mention here some of these efforts<sup>2</sup> such as Mika [49], Gruber [97], Halpin et. al. [96], Cattuto et. al. [98], Borwankar [99], Story [100], Newman [101], Knerr [43], Passant et. al. [102], Scerri et. al. [103], and Kim et. al. [104].

### **Using ontologies in the folksonomy**

In this category of methods, the Social Web exploits the Semantic Web's structured vocabulary by consulting the controlled vocabulary of ontologies to extract any relations that add meanings to the folksonomies' tags. The consulted ontologies might be domain ontologies or lexical ontologies.

1. *FolksAnnotation method* [4]: This method generates semantic metadata for learning resources by using folksonomies guided by domain ontologies. The system has two stages as shown in Figure 3.2. In the first stage, all tags assigned to a learning resource in del.icio.us are extracted and normalised to clean up the noise in people's tags. The normalisation process includes converting tag to lower case, removing non-English tags, stemming, grouping similar tags, and eliminating general tags. The second stage is the semantic metadata creation where all the normalised tags are adhered to different domain ontologies' concepts, and only the terms that appear in the ontologies will be selected as "semantic metadata".

---

<sup>2</sup>For more details, we refer you to the associated references.

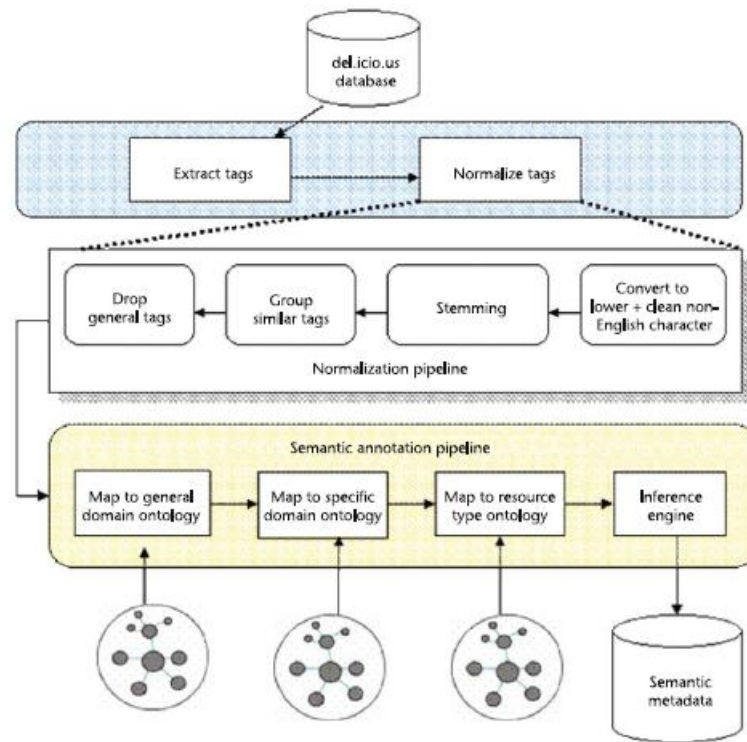


Figure 3.2: FolksAnnotate architecture [4].

This method is limited to specific domain(s). It modifies, and even eliminates, many user tags. Furthermore, the evaluation results rendered in this paper were preliminary and not enough to prove the validity of this method.

2. *Folksonomies and Ontologies in Authoring of Adaptive Hypermedia* [5]: This method combines folksonomies with ontologies to create semantic relations among the folksonomy's tags. The merged methodology of Web 2.0 and Web 3.0, as shown in Figure 3.3, consists of three phases; filtering, grouping, and mapping. In the *filtering* phase, the Google spell checker software was used to replace the misspelled tags with the suggested correct ones. In the *grouping* phase, the similar tags are grouped together based on their mutual co-occurrence values. These two phases occur at the Social Web side. Although grouping tags is a first important step, it does not give any information about the structure of relations amongst these tags. Enriching the tags in each group with semantic relations occurs in the *mapping*

phase which occurs at the Semantic Web side. In this phase, the online ontologies are found using the semantic search engine Swoogle in order to achieve the mapping process between grouped tags and elements of matching ontologies. Based on the tags co-occurrence and the hierarchy of the matching tags in the relevant ontologies, this phase generates a hierarchy that represents a bottom-up ontology from folksonomies rather than a predefined ontology.

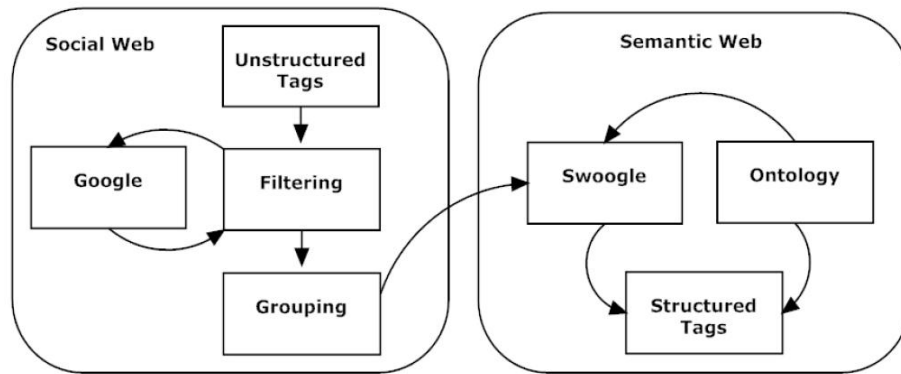


Figure 3.3: Merging Social Web with Semantic Web [5].

This hierarchy can be used for auto-replacement of a particular string with another one while tagging, such as “NYC” (unstructured tag from folksonomies) would be replaced with “New York City” (structured tag taken from the *city ontology*). Auto-completing of tags while tagging is another use of the produced bottom-up ontology.

This method builds an ontology that mixes the controlled vocabulary of ontologies with the free tags of folksonomy to produce a bottom-up ontology that will be used for auto-replacement and auto-completing of tags. From a user point of view, we argue that the interference in user tags by replacing them with other tags is not acceptable. Moreover, the auto-completion may seem convenient for some users, while it is not for others.

3. *Using WordNet to Turn a Folksonomy into a Hierarchy of Concepts* [6]: This method integrates an ontology in the interface of a folksonomy to add some ex-

explicit semantics provided by a static hierarchy of concepts; the chosen folksonomy in this method was del.icio.us, and the ontology was WordNet. In particular, this method used the WordNet concepts' relations to show the user an additional panel on the browser's interface; it is the *tags semantic tree* panel (see Figure 3.4). This extra visualisation displays a higher number of related tags organised according to semantic criteria.



Figure 3.4: A screenshot from the Del.icio.us page for tag “Pasta” - the inner sidebar shows an expandable hierarchy of related tags. [6].

The aim of this method was to guide the user by displaying more related tags that will facilitate navigation and searching in the folksonomy, and to support the pro-



cess of semantic tagging; it is all about visualisation.

4. *OntoSonomy* [7]: The main idea of *OntoSonomy* is to give meanings to the tags by combining folksonomic tagging and ontology. It considers using both domain-specific ontology and generic ontology (WordNet) when annotating, searching, and browsing in the prototype system. According to statistical studies, the most popular tags used in the folksonomy were specified, and then the domains of these popular tags were determined. For example; the most popular tags lie in the travel domain. Once the domains are specified, an ontology for that domain is manually built; so-called the domain-specific ontology. The user tags and their related tags extracted from the WordNet are then filled as instances in the defined ontology. Consequently, while tagging, the user needs to provide classified tags in different text fields, similar to filling in a form, to match the built domain ontology as shown in Figure 3.5. Searching the *OntoSonomy* is done in a similar way.

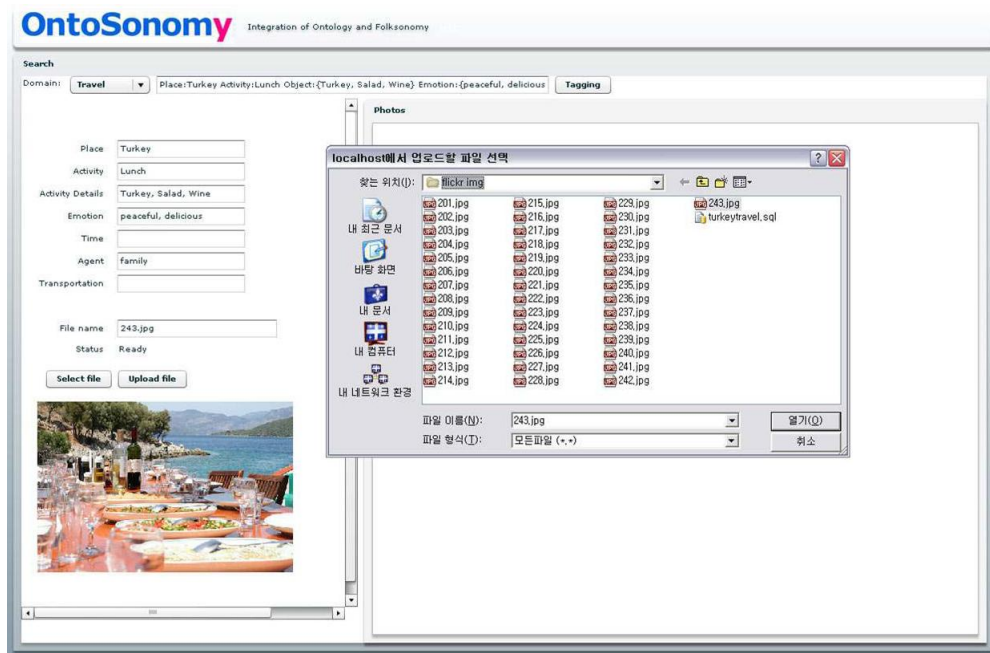


Figure 3.5: *OntoSonomy* prototype interface [7].

Obviously, this method is changing the conventional interaction pattern between the user and the folksonomy. It suffers from the absence of well-known simplicity

of tagging systems. Moreover, manual building of domain ontologies is a prerequisite in this method. So, it will be always demanding to build emerging domains' ontologies.

5. *TagOnto method* [105]: TagOnto is a folksonomy aggregator to bridge the gap between Social Web and Semantic Web by automatically mapping the unstructured tags to more structured domain ontologies. It provides ontology-based searching capabilities to combine results from different tag-based systems (e.g. Flickr, YouTube, etc). Once the user searches for a tag, the matching concepts of that tag are generated from the associated domain ontology. If the tag appears in the domain ontology with different senses, the system can select the suitable meaning. The disambiguation process is performed in two steps; retrieving the most frequent co-occurring tags to define a context for the wanted tag, and then analysing the ontology to discover the meaning of the tag in that particular context. As the searching tag is disambiguated, the results will be retrieved from many tagging systems in one screen.

Obviously, this method tries to address the polysemy problem while searching the tagging systems. The experiment of this method was done using the “*Wine Ontology*”. We argue that applying this algorithm in folksonomies will need unlimited domain ontologies to cover the various tags that users use in their search. Building these ontologies is time, and effort, consuming.

6. *Relating User Tags to Ontological Information method* [106]: This method aims at applying both syntactic and semantic techniques for connecting a tag to ontologies in order to get more semantics about the tag. It develops a reusable generic component called “*Matching Component*” that can be used with any folksonomy. This component takes the user tag(s) as input to generate matching, as well as semantically related, words in the WordNet. The generated tags are provided to the user as

a set of suggestions. The matching component has a feedback channel that considers two types of users feedback; *implicit* and *explicit* user feedback. The implicit feedback is automatically obtained by the system by recognising which suggested tags were chosen by the user, while the explicit feedback is asking the user which suggestions were good and which were bad.

This method, in addition to offering suggestions to the users, it asks the users to give feedback about these suggestions. Hence, we argue that it puts more effort on the users' side to improve the quality of the tags by changing the conventional way by which the users used to interact with the folksonomy.

### 3.3.2 Social networks approach

The social networks methods are based on the tripartite of the folksonomies themselves; tags, taggers, and tagged resources. In other words; the folksonomies use the folksonomies to solve the problems of folksonomies. These methods illustrate and analyse the relations not only within each element of the tripartite mentioned, but also they focus on the interrelations among the actors, resources, and tags [44].

1. *Automated Tag Clustering method* [8, 44]: Data clustering is a statistical technique for data analysis that groups the whole dataset into similar subsets; these smaller subsets of data are called clusters [8, 107].

The automated tag clustering method deals with the search problem in folksonomies. The problem comes from the fact that different users use different tags for the same content. The problem is easier when huge number of users annotate the same object; because this creates a kind of users' consensus on their tags. The situation becomes worse when only few users annotate one object with high diversity in their tags. The automated tag clustering method claimed a quite good solution for the latter situation. The algorithm builds clusters of tags in the folksonomy; each clus-

ter contains exact tag, related tags, and a weight which is the number of times in which the related tag co-occurred with the exact tag in that folksonomy. Figure 3.6 exemplifies that the tag *design* co-occurred 2084 times together with the tag *Web* in the same folksonomy, and it co-occurred 728 times with the tag *inspiration*.

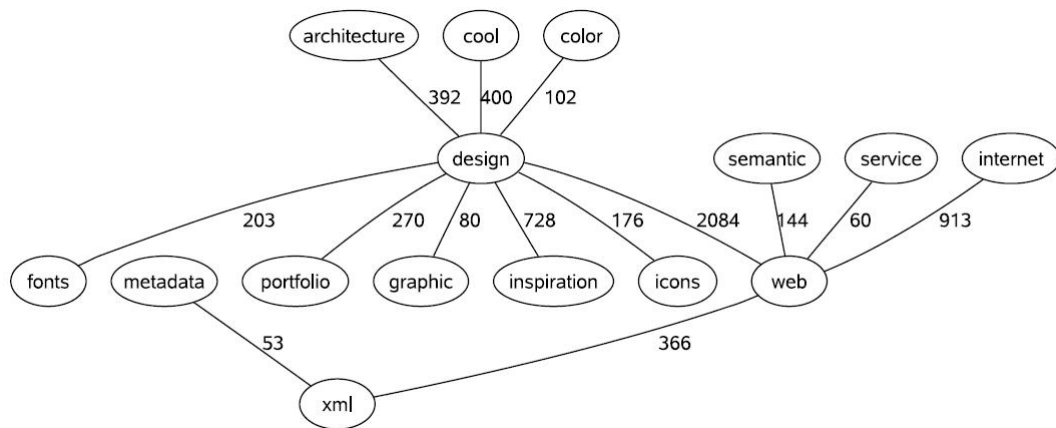


Figure 3.6: Part of the cluster for the tag “design” [8].

The algorithm selects the top  $N$  related tags and stores them in the folksonomy database as shown in Table 3.1.

Tag	Co-tags (related tags)
Design	web, inspiration, cool, architecture
Web	design, internet, XML, semantic
Apple	mac, osx, macosx, tiger
Art	cool, design, fun, graphics
Javascript	ajax, dhtml, programming languages
Photography	galleries, photo, hi-res, sexy
Music	audio, media, mp3, ipod

Table 3.1: Example of *Top-4* related tags for some tags [8].

To illustrate the use of these clusters in search, assume that a page on del.icio.us was tagged as (*apple, mac, osx, car*). In the clusters already obtained for del.icio.us, there was no cluster that says *car* is a co-tag (related tag) for the other tags. At the

same time, there are clusters that say the other tags are co-tags. Then the system will know that the tag *car* is an *odd* tag for that page. And thus, that page will not be retrieved when the user is searching using the *odd* tag *car*.

The dynamicity of this method is that the clusters are inferred from analysing the patterns of users' behaviour in tagging, and they are not pre-determined constant data. Moreover, these clusters give a great *context* to define tags.

2. *Tag Contextualisation method* [78]: This method shows the role of the social context and how, when considered, it gives a better picture of semantics of tags without consulting any external resources. Clustering the same group of tags many times based on different criteria obtains a set of clusters for the different contexts in which an ambiguous tag is used. Comparing a very small set of results (only 10 tags) with WordNet reveals some interesting facts; on one hand the clustering method does not give all the meanings of a tag retrieved from WordNet, on the other hand WordNet does not contain all the useful meanings obtained from the folksonomy. Therefore, the social knowledge existing in folksonomies can work together with the controlled vocabulary offered by Semantic Web ontologies to leverage the semantics in social networks.
3. *Community Based Folksonomy method* [69]: This method proposes the exploitation of the metadata existing in wide spreading social networks. It maps the tags of the users in one tagging system with their friends' tags in other tag-based systems on a small community basis; this is so-called matchmaker-based recommendation system. The system allows the users to add bookmarks, tag them, and browse the friends' bookmarks to map their tags with their friends' tags. Personal link between the user and the user's friends can be done by using the FOAF ontology. These functions are managed by the three main components of this system; *personal contents manager* (for adding bookmarks and tags), *personal network manager* (for tag surf-

ing and social networking), and *contents recommendation manager* (for suggesting the tags used by friends).

We argue that grouping the users into smaller similar groups will complicate the aforementioned problem of *specialised tags*. Moreover, it complicates the existing conventional tagging behaviour.

4. *Marlow et al. method* [9, 44]: The research team suggested a conceptual model for Web-based tagging systems (with no empirical results). This method focused on the internal analysis of the folksonomy itself; the relations among the tripartite *tag-user-resource*. As seen in the conceptual model in Figure 3.7, not only the tags connect the users with the resources, but also similar resources may be connected to each other, and users that have common interests may also be connected to each other. By considering such a model, the researchers claim that there is a possibility to infer some semantics by segmenting the structure of the social network. For example, when some portions of users use certain tags for the same resource, or correlated resources, this may imply that these tags are synonyms.

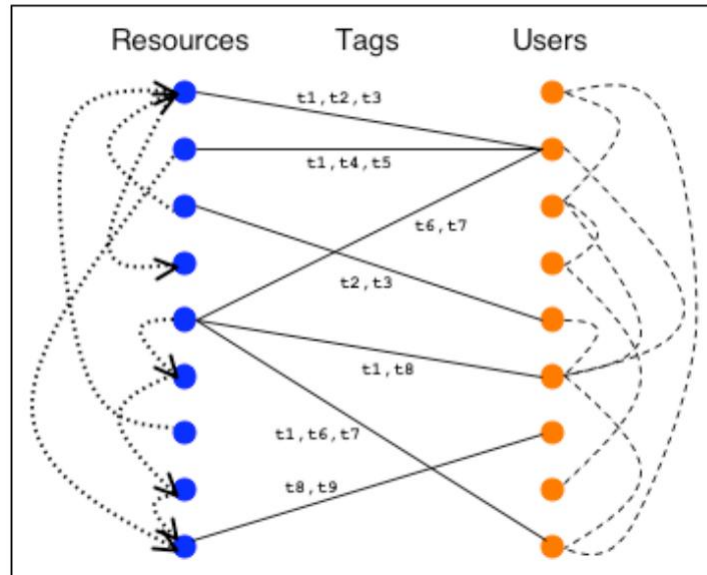


Figure 3.7: A model of tagging system [9].

### 3.3.3 Visualisation Approach

Improving the users' behaviour could play a considerable role in improving the tagging process. One of the approaches used to increase the users' awareness in tagging is the visualisation approach. This approach involves any visual aids used to display tagging information (tag-user-resource) to the end-users. Usually this visualisation is built on the analysis of users' tagging behaviours and patterns.

1. *Improving Tag Clouds method* [10]: The tag cloud visual model is well known and widely used in folksonomies. The selected set of tags to display in such tag clouds is the most frequently used tags. The tags in tag clouds are arranged in alphabetical order, which does not facilitate visual scanning or discovery of semantic relations among tags. This method proposes new presentation of tag clouds where similar tags are grouped based on co-occurrence analysis to improve browsing experience. The improved tag cloud, shown in Figure 3.8, arranges the displayed tags together according to different meaningful criteria; the semantically similar tags are horizontal neighbours in the same cluster, and likewise, the semantically similar clusters are vertical neighbours.



Figure 3.8: Improved tag cloud [10].

We advocate this new approach as alternative of traditional tag clouds since it contributes in improving the visual consistency of represented tags. It adds semantics to the browsing experience.

2. *Cloudalicious method* [108]: Cloudalicious is an online visualisation tool that shows the growth and changes of the tag clouds over time. It is available online to all users; any user can visit <http://cloudalicio.us/tagcloud.php> at any time to visualise the tags used in del.icio.us for a given Uniform Resource Locator (URL). The website asks the user for a URL, downloads the tagging data from del.icio.us, then plots the users tagging activity over time. Figure 3.9 shows an example generated by <http://cloudalicio.us/tagcloud.php>; the x-axis shows the interval when the tagging data was taken, while the y-axis shows the weight (weight = times used / number of authors) of the most popular tags for that URL. The longer time the line moves from left to right, the more stability there is in the tagging activity.

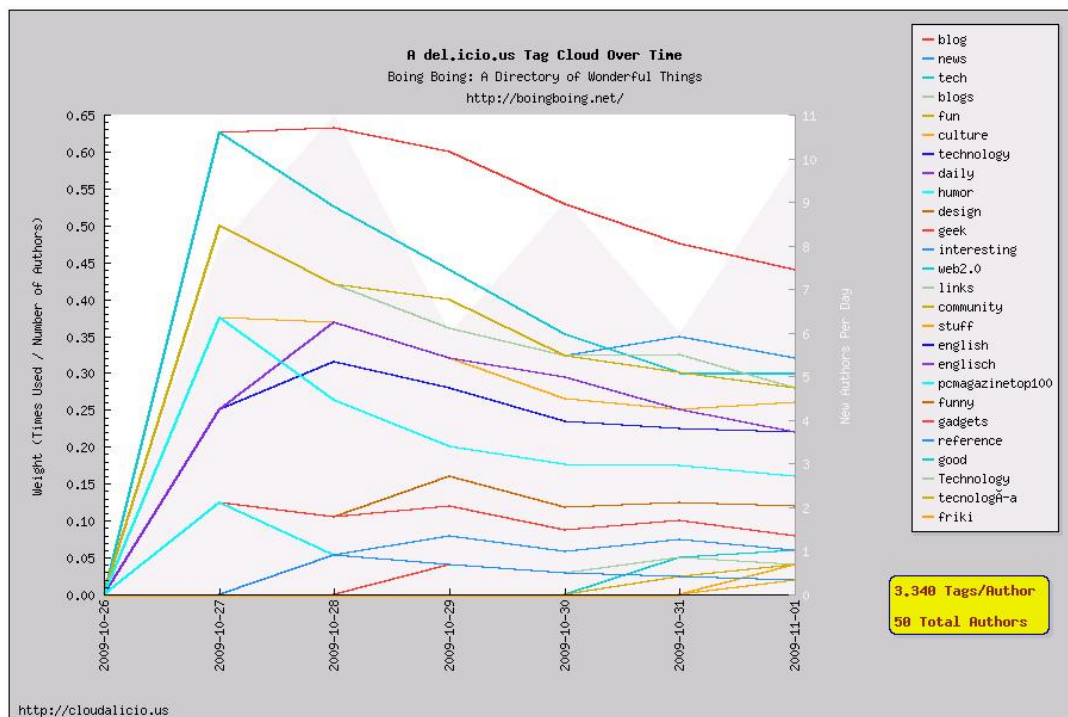


Figure 3.9: A Del.icio.us visualisation of A tag cloud for Boing Boing website generated by <http://cloudalicio.us> for the interval (26-10-2009 to 01-11-2009).



This visualisation tool is not available in the folksonomy at tagging time. Therefore, it does not help in improving the tagging activities. This tool is more appropriate for the savvy-tech users who show high interest in the tagging and they are keen to learn how to create more accurate and consistent metadata. We argue that the percentage of such users is relatively low. On the other hand, Cloudalicious presents a free tool for researchers to quickly identify patterns in taggers' behaviour in a timely fashion.

### 3.4 Summary

Several statistical studies have been conducted to explore the regularity, stability, and users' consensus in tagging activities. Such studies are prerequisite for attempting to address any of the folksonomies shortfalls.

By classifying the research efforts in addressing the drawbacks in social tagging, we could identify three main approaches; ontological approach, social networks approach, and visualisation approach.

The *ontological approach* comprises the methods that tried to use the power of the controlled structure of the Semantic Web ontologies. This approach shows the best results in addressing the problems of ambiguous and inconsistent classification of data in folksonomies [44].

The *social approach* showed that one folksonomy can learn from the social knowledge existing in other folksonomies and exploit the diversity of data sets they have. Moreover, despite the lack of controlled classification of data in folksonomies, there still are valuable and trustable semantics due to the regular patterns in tagging.

The *visualisation approach* might facilitate the users' browsing experience in social tagging websites. Nevertheless, it does not address the core problems in folksonomies; ambiguity and inconsistency. Thus, we limited ourselves to mention some of these methods to illustrate the approach.

None of the related works covered in this chapter could address the challenges of semantic relations, multilinguality, and shorthand writing. Moreover, there are search trials that tried to address some tagging systems challenges in one hand, but on the other hand they either alter the users' tags or violate the simplicity of tagging process. Furthermore, some solutions proposed in the related works showed incompatibility with current tagging systems.

## **Part II**

# **Tagging System Architecture**

## Chapter 4

# Generic Architecture for Tagging Systems

### *Objectives:*

---

- Introducing general criteria for approaches addressing the tagging systems challenges.
  - Providing a generic architecture for addressing majority of tagging systems challenges.
  - Introducing the idea of adding “*system tags*” based on the “*user tags*”.
-

## 4.1 Introduction

Having reviewed the various techniques of tagging systems, studying their natures, advantages, disadvantages, and challenges, and after investigating the current attempts to improve tagging systems efficiency, we have come up with some criteria that we perceive as crucial in leveraging tag-based systems.

Keeping an eye on these criteria and the challenges of tagging systems, we have produced a generic architecture that can address most of the tagging systems challenges. This architecture comprises semantic, as well as social, aspects.

## 4.2 Standards and criteria for an efficient approach in developing a tagging system

There are a number of important rules that should be adhered to when designing new approaches for tagging systems. These are summarised as follows:

### 4.2.1 Integrity of user tags

Some trials for normalising tags have intended to remove unwanted tags (see [4]). *Integrity of user's tags* means avoiding any deletion of user tags even though some tags seem to be noisy. The thesaurus of tag-based systems has been constructed over a period of time. Therefore, any tag which might look strange, or noisy, today will have a meaning in the future. Vice versa, what may be considered as known tag today might have been considered strange some time before. Users invent new terminology or add new meanings to old words. So a tag which appears meaningless to the system, or to most users, might be a neologism which becomes popular over time. Users have meanings for tags in their

minds while tagging. Instead of removing strange tags from the tagging system because they are not understood, an effort should be paid to define, or even, to clarify these tags' meaning.

Tag elimination threatens the construction and the evolution of the tagging system thesaurus over time. Any tag offered by any user is highly respected; the irrelevant tags are as valuable as the relevant ones. The user tags are of high importance since they reflect the emerging vocabulary in the social communities. Moreover, they represent the social asset or, more likely, the social intelligence.

Furthermore, even changing or updating the user's tags is not accepted; users will be dissatisfied if they add some tags and the next day they discover that the system is changing these tags (see [5]). Forms of unaccepted changes are, for example, for the sake of tags uniformity such as converting plural form to singular forms, or converting all tags to lowercase. Such processing can be done internally for computational purposes but the users should not see it.

*Do not interfere with the user tags*

### **4.2.2 Integrity of social interaction patterns**

Social websites have introduced new patterns of interactions between the users and the Internet which were not existing before. Furthermore, the Social Web has created a platform to extend the interaction to be among the users themselves rather than just between them and the Internet. Such emergence of e-socialisation has gained unexpected success that excited the individuals and organisations in the Web community.

Accompanied with this trend, there have been some important features that attracted

the users to be part of the process of producing and consuming content in addition to the networking phenomenon. The interaction pattern between the users and the Web was one of the main reasons behind the involvement of vast numbers of users in this trend. The characteristics of such pattern are summarised in the following:

- *Simplicity*: simple and clear functionalities where any user, with very minimum knowledge of computer and information technologies, can understand and interact with the web. User-friendly Web design plays a main role in simplicity.
- *Ease of use*: because it is simple, it is easy to use.
- *Informal platform*: the core idea in the Social Web is that users can express themselves using whatever language, terminologies, jargon, pidgin, slang, or colloquial speech they want; it is a platform where everybody, evenly, has a voice.

Tagging systems, which is in the heart of the Social Web, are so popular, widely spread, and accepted because they conform with the aforesaid characteristics.

As a trial to make some improvements in tag-based systems (e.g. data classification, tag normalisation, information retrieval, adding semantics, etc), some researchers have invented new methods that make improvements on one hand, but contradict with the ethos of the current interaction patterns on the other hand (see [81]). *Integrity of social interaction patterns* means avoiding radical changes in these patterns since they have proven their wide success and acceptance, and have been behind the popularity of tagging systems.

New attempts to improve the efficiency of the tagging systems must not change the way in which users create and share their tags. Putting new effort on the user side is not a wise decision. The need is to address the challenges of the current tagging systems while

keeping the current behavioural interaction pattern.

*Do not change the user's interaction pattern*

### 4.2.3 Rich functionality

As mentioned before, all tagging systems share the same aforementioned challenges. These systems exist in reality and are successful with some drawbacks. That means; the need is not to create new tagging systems, rather, it is to create new solutions for the current ones. A brilliant solution is the one that comes up with an architecture which is applicable in the current systems. The proposed architecture should act as a toolkit for any tagging systems, the current tagging systems and the potential future ones, to enrich their functionalities.

In other words, tagging systems need adding some new features in the system background, not in the foreground. The users will not see these features but they will feel their consequence and significance in the outcomes (e.g. search accuracy).

*Create applicable/compatible function for the current systems, not  
a new entire system*

### 4.2.4 Universality

Tagging systems, or even all Social Web applications, target everybody; the target community is the world community with no limits. A robust architecture for such tagging applications should expect users from different classes. Tag-based systems users belong to various generations coming from different places over the world, and having different social, educational, and cultural backgrounds.



Targeting a niche group of users contradicts with the *universality* of e-socialisation. While judging a new approach for addressing tagging systems' challenges, it should be checked whether it is dedicated for a special group of users or it targets the world community (see [39]). Any segmentation would be undesirable no matter the criteria considered for grouping the users.

*Bearing in mind that tagging system targets everybody, do not dedicate it for a special group of users*

### 4.2.5 Dynamism

Tags in tagging systems are dynamic and trendy by nature; new meanings of the vocabulary used in tagging appear from time to time according to the changes in the real social communities to which taggers belong. Nevertheless, some tags have static meanings over the time. This dynamism in tagging systems is an inevitable reflection of the dynamism in natural languages; tagging systems are the mirror of the people's slangs.

In order to add semantics for the emergent trends in the language used in tagging, tagging systems should keep a dynamic resource on which they depend to discover the semantics for such unexpected trendy tags. Yet, static resources are necessary all along the way.

*Make sure that the tagging system has dynamic resources, in addition to the static ones, to adapt with emergent trends*

### 4.2.6 Multilinguality

As aforesaid, universality is a main trait of tagging systems; they are available online and one click away from users that come from different countries, and thus, speak different languages. With no doubt, an exact tagger for a specific content (e.g. video on YouTube

or bookmark on Del.icio.us) will tag using a language that is not understandable by all potential future searchers for that content. As a result, that content will not be retrievable using any language, but the exact tagger's language.

The tagging system should provide a mechanism to add language-independent accessibility to its content. This will make the content survivable and retrievable for cross-languages users.

*Make the tagging system speak different languages*

### 4.3 Generic architecture for tag-based systems

Having investigated tagging systems and being aware of their nature, notion, principles, characteristics, advantages, and challenges, we built a generic architecture for tagging systems. The architecture keeps the ethos of tagging systems and conforms to the rules abovementioned.

This architecture can be considered as a template for improving the efficiency of tagging systems on one hand, and on the other hand, it assures that the essence of tagging systems is safeguarded. Therefore, future trials to address the challenges of tagging systems should keep an eye on the components of this architecture.

The main aspects of the improvements in this architecture are; the addition of semantic dimension, multilinguality, and clustering.

Figure 4.1 shows the outline of our architecture. The prototype (will be discussed in more details in the following chapters) will be built as depicted this architecture.

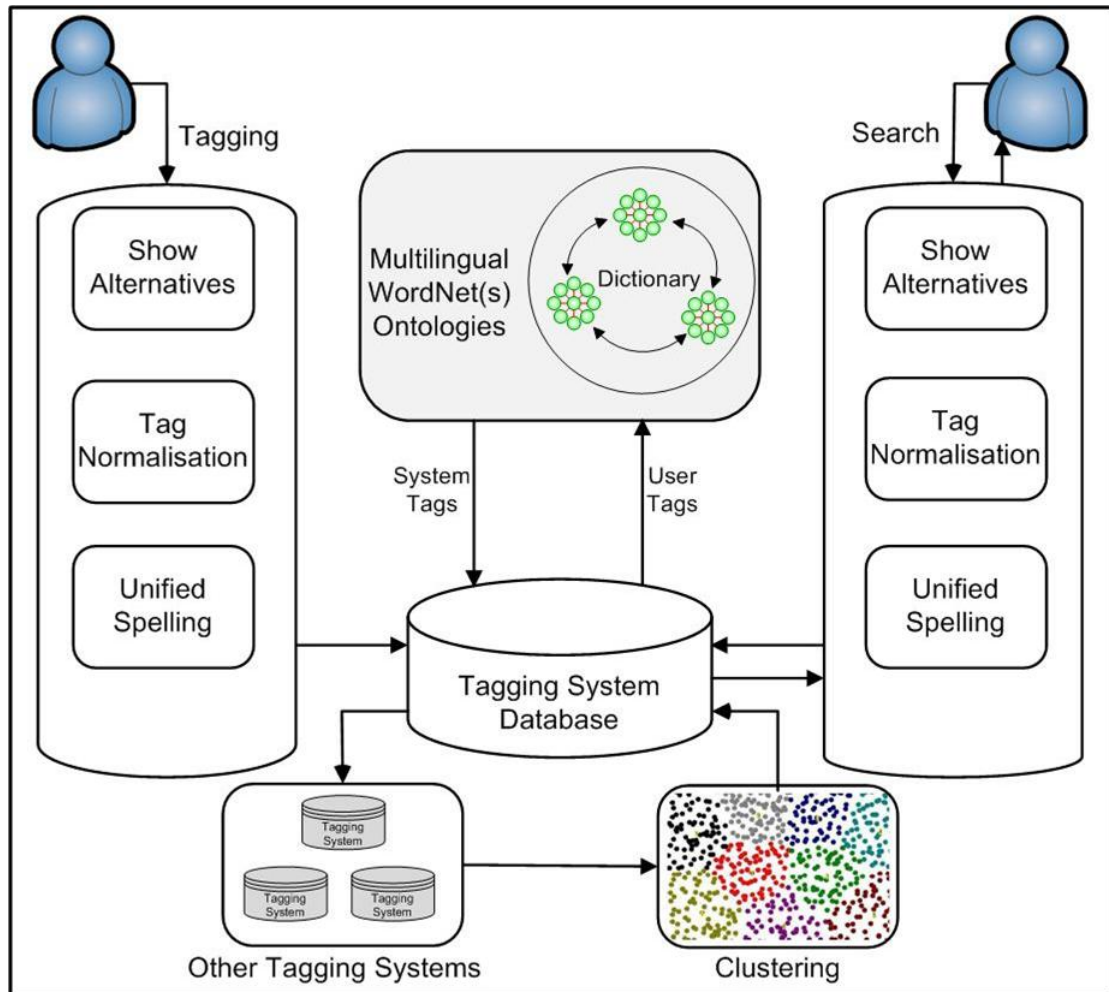


Figure 4.1: Generic architecture for tag-based systems.

### The architecture's use case diagram

The first interaction form between the user and the tagging system takes place when the user uploads new content to the tagging system portal (e.g. photo to Flickr). It is not mandatory in most tagging systems to tag the uploaded content; adding tags is optional. This is due to the fact that these systems do not depend only on tags as metadata, but also on other textual metadata such as title, description, user name, etc.

In our work, and in the context of tagging, we consider that tagging is the first interaction form between the user and the tag-based system.

In the *use case* diagram of our architecture shown in Figure 4.2, we present a graphical overview that illustrates the behaviour of the tagging system and the role played by its actors (users).

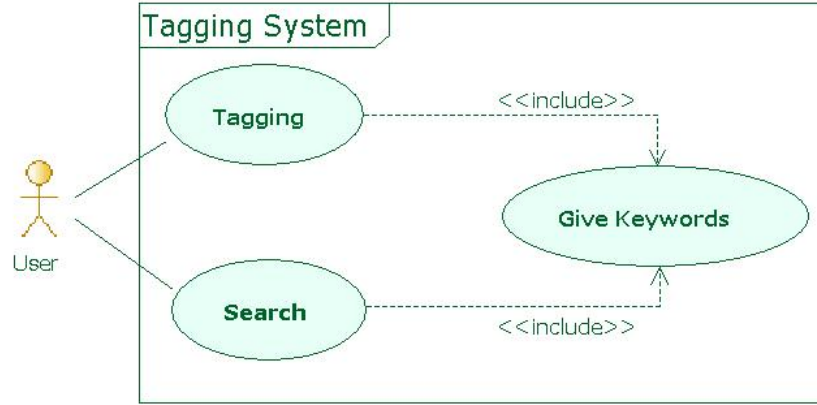


Figure 4.2: Tagging system use case diagram.

The interaction between the users and the tagging system takes place either by *tagging use case* or by *searching use case*<sup>1</sup>. Definitely, both of *tagging* and *searching* use cases imply (<<include>>) inserting some keywords; these keywords are *tags* in the former use case whilst they are *search terms* in the latter use case.

### Challenges to be addressed

Since we are trying to address the tagging systems challenges in this architecture, we briefly mention these challenges as a reminder for readers. The first nine challenges were already discussed in the literature review. The last two challenges (*shorthand writing* and *semantic relations*) have not been mentioned before in the literature. Therefore, we discuss them in this section. The challenges are:

1. Word synonyms
2. Word polysemy (homonym)

<sup>1</sup>More details will be spelt out in 4.3.1 and 4.3.2.

3. Different lexical forms
4. Alternative spellings
5. Misspelling errors
6. Badly encoded tags
7. Specialised tags
8. Key phrases instead of keywords
9. Different languages (multilinguality)
10. Shorthand writing
11. Semantic relations

### **Shorthand writing tags**

By exploring the Social Web nowadays, it is notable that users have started using *short forms* of some words similar to, and might be influenced by, the language used in Short Text Messages (SMSs) on mobile phones. These forms are not recognisable as words in the lexicon [6]. Table 4.1 shows some examples of these short forms.

This trend might be correlated to the appearance of special abbreviations used in SMSs. Moreover, it might be correlated to the new tools emerged for accessing the Internet such as mobile phones.

With the advances in mobile technologies, new mobile generations have been launched that enable the users to explore the Internet via their mobile devices. Users nowadays are using their mobiles increasingly to socialise among each other by visiting online social websites. The majority of mobile keyboards do not facilitate the typing process at the

Complete Form	Short Form
text	txt
love	luv
good	gd
night	nite
mate	m8
great	gr8
later	l8r
before	b4
tomorrow	2mro
oh my god	omg
by the way	btw
be right back	brb

Table 4.1: Short forms vs. complete form of some English words/phrases.

same level of easiness as computer keyboards do.

We argue that the use of *short forms* of words, we name it “*shorthand writing*”, is a crucial challenge to consider in the context of tagging systems.

### Semantic relations

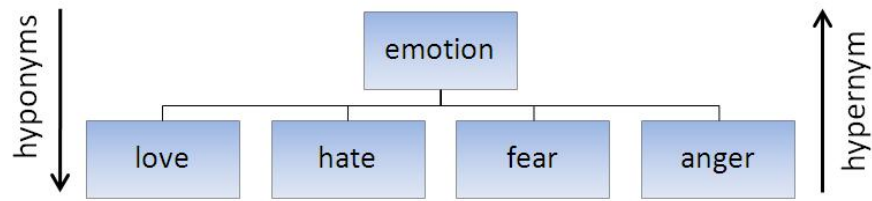
In the literature, all resources spotlight synonymy as a challenge of tagging systems. Indeed, synonymy is one kind of semantic relations between words whereas there are many other relations such as hypernymy/hyponymy, meronymy, etc. In other words, the “*semantic relations*” is the superordinate whereas the “*synonymy*” is the subordinate.

We deem that the challenge is broader than just one kind of semantic relations (synonymy); synonymy is just part of the scene. This part comes into view in tagging systems when a user searches a tagging system using a specific keyword. The results that will be retrieved are all the objects in the tagging system that were originally tagged using that

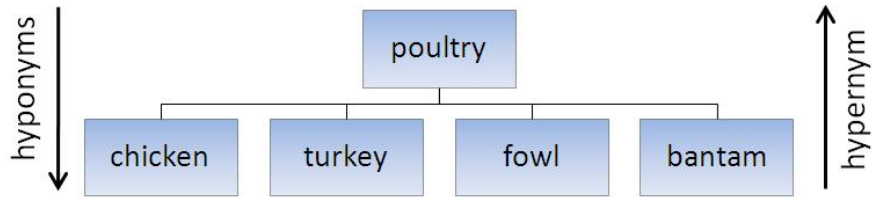
specific keyword, but not the objects that were tagged using the synonyms of that specific keyword. For example, when searching using the keyword “*cellphone*”, the result will be the objects that were originally tagged using the word “*cellphone*”, but not the objects that were tagged using any of the synonyms “*cellular telephone*”, “*cellular phone*”, or “*mobile phone*”. Even though the objects that were tagged using the synonyms of that specific keyword are related to that specific keyword and expected by the user to be retrieved, unfortunately they will not be retrieved.

We argue that the same scenario is happening in some other kinds of semantic relations; not only in synonymy case. In hyponymy, for example, when a user searches a tagging system using a specific keyword, it is not expected to retrieve only the objects that were originally tagged using that specific keyword, but also the objects that were tagged using the hyponyms of that specific keyword. For example, when searching using the keyword “*emotion*”, the result will be the objects that were originally tagged using the word “*emotion*”, but not the objects that were tagged using any of the hyponyms “*love*”, “*hate*”, “*anger*”, “*fear*”, etc. Another example is when the used search keyword is “*poultry*” the objects that were tagged using any of its hyponyms (“*chicken*”, “*turkey*”, “*fowl*”, “*bantam*”, etc) will not be retrieved although they are related and expected to be retrieved. These two examples are illustrated in Figure 4.3. If the relation is read top-down, it represents a hyponymy relation whilst it represents a hypernymy relation if it is read bottom-up. Further discussion about the semantic relations will come in the next chapter.

Accordingly, tagging systems do not consider the semantic relations between the saved tags and the search terms; tagging systems can read the natural languages but cannot understand them yet.



(a) The hypernym “*emotion*” and some of its hyponyms.



(b) The hypernym “*poultry*” and some of its hyponyms.

Figure 4.3: Examples of hypernymy/hyponymy semantic relations

From the examples above, it is obvious that the “*semantic relations*” challenge is the superior challenge in tagging systems.

Along with the illustration of our architecture, we keep one eye on these challenges to check which were addressed by the architecture. At the same time, we keep the other eye on the abovementioned rules, if there is any relevance.

In order to describe our architecture, we divide it into five components; tagging component, searching component, semantic component, clustering component, and database component.

### 4.3.1 Tagging component

*Tagging component* is the highlighted area shown on the left hand side of our architecture in Figure 4.4.



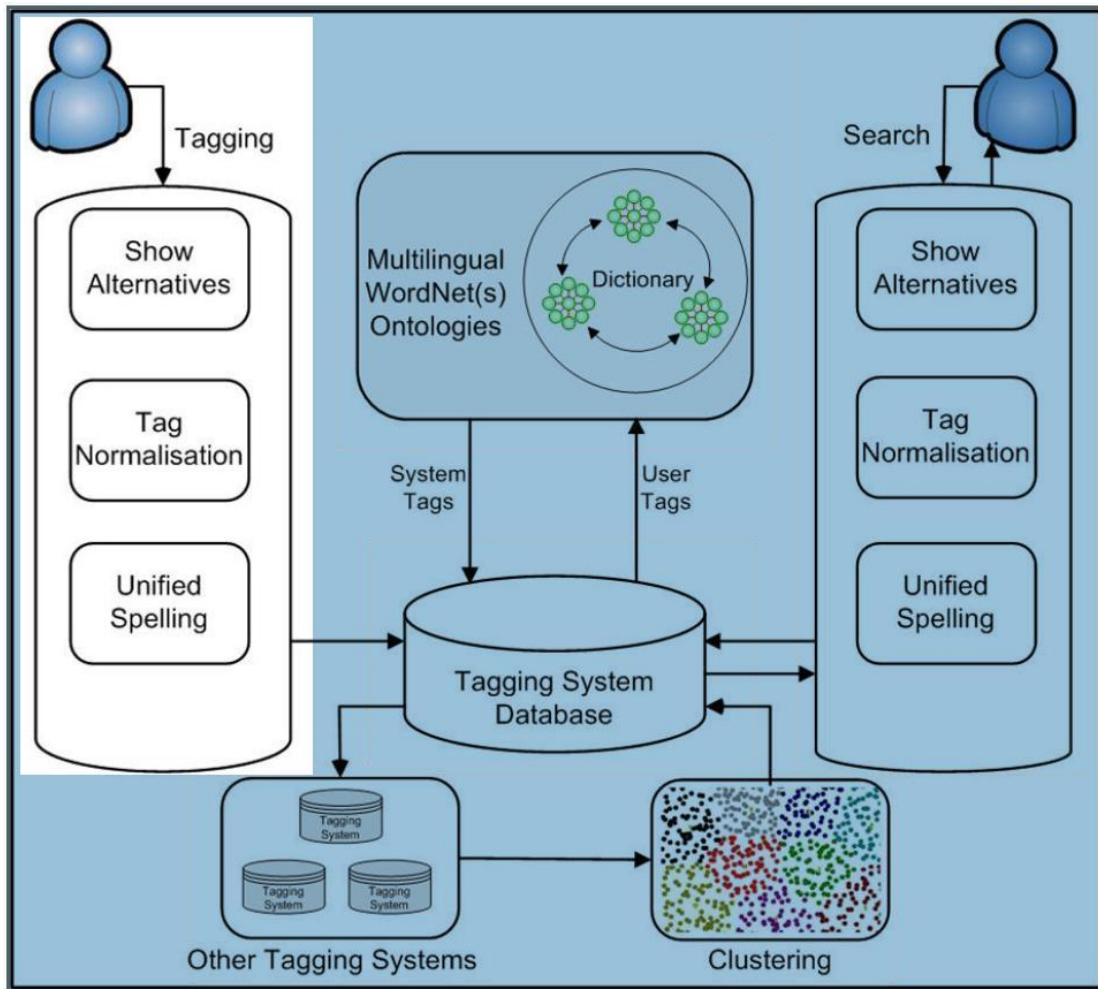


Figure 4.4: Generic architecture for tag-based systems - *Tagging component*.

Since tagging is a form of interaction between the user and the tagging system, it is important to make sure that the well-known patterns of user-system interactions are not changed. With respect to the aforementioned “*integrity of social interaction patterns*” criterion, users should not be enforced to accomplish any extra functionalities such as enforcing them to use some keywords alternatives. Furthermore, the way in which they interact with the system interface should be kept as conventional as possible.

When tagging, users are just giving free-text keywords that are best describing the tagged object. The system, as seen in Figure 4.4, will perform three processes; showing

alternatives, normalising the user tags, and unifying tags spelling. The activity diagram of *tagging component* is shown in Figure 4.5.

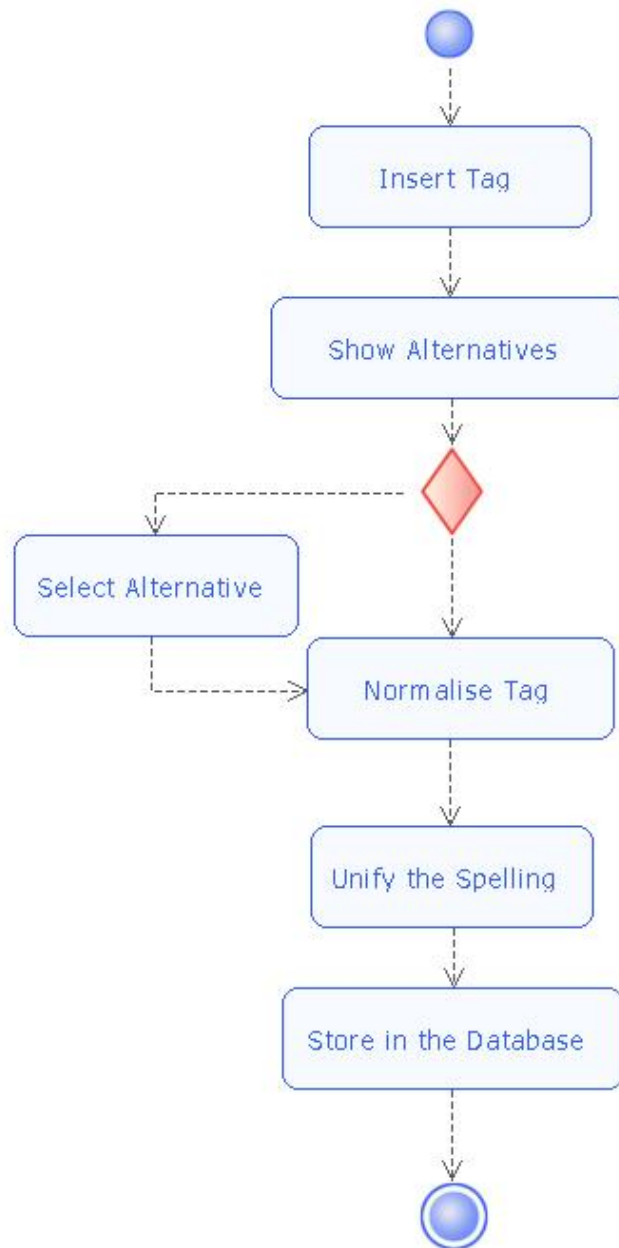


Figure 4.5: Tagging activity diagram.

1. *Showing alternatives*: Once the user has inserted a new tag, the system will show some alternatives for the inserted tag. Alternatives are the corrected forms of the word if there are spelling errors or badly encoded tags. The user has the choice ei-

ther to replace the inserted tag with the suggested one or to ignore any suggestions and continue the conventional way of tagging. The way in which the system shows these alternatives should not affect the user's typing process. If the user decides to select the suggested alternative tag, (s)he can scroll to the suggestion using the input device (e.g. keyboard, mouse, touch screen, etc).

To show the corrected words as alternatives, a *spell checker* application can be used similar to those used in word processors, email clients, or search engines.

By showing optional alternatives, the architecture is addressing the challenges of *misspelling errors* and *badly encoded tags*. At the same time, the *integrity of social interaction patterns* criterion is not violated. Yet, some misspelled or badly encoded tags will still exist in case the user ignores the suggested alternatives.

2. *Normalising tags*: Whether the user selected a suggestion or not, the next process for the tag is the tag normalisation. Normalisation means the reverse process of the different lexical forms of a word to the original lexical form. Consequently, all plural nouns will be reverted to the singular form (*men* will be *man*), and conjugated verbs will be reverted to their bases (*ate*, *eaten*, or *eating* will be *eat*).

With respect to the *integrity of user's tags* criterion, this reversion of tags should not interfere or change the user's tags. Furthermore, the user should not be able to see this process because of the *integrity of social interaction patterns* criterion. To obey the rules of the two criteria, the tagging system should keep two copies of tags; the tags inserted by the users without any change and the corresponding normalised tags. The purpose of keeping a copy without any change is that it will be used for the future display by the taggers themselves so that they will not be aware of any

change of their original tags. Besides, this copy might be used as *raw metadata* for future searching by other users. The normalised copy will be used intentionally as *normalised metadata* for future searching by other users, as well as, for further semantic processing after the tags are stored in the tagging system database<sup>2</sup>.

To revert the conjugated words to their stems, some existing *stemming algorithms* can be employed by the tagging system (e.g. Krovetz algorithm, Dawson algorithm, Porter algorithm, etc [109]).

By normalising the tags, our architecture is addressing the challenge of using *different lexical forms*. At the same time, the normalisation process complies with the *integrity of user's tags* and *integrity of social interaction patterns* criteria.

3. *Unifying tags spelling*: “Color” and “colour” are, respectively, the American and the British spelling for the same English word. To avoid considering these two spellings as two different tags, we consider one unified spelling to store in the database. The number of different spellings in other languages, such as Arabic language, might exceed two spellings for the same word.

Therefore, after the tag is normalised, one unified spelling will be considered (e.g. British English). Likewise in the tag normalisation process, two copies will be stored in the system for the same reasons abovementioned.

Considering one unified spelling, one might need to build a small database where all different spellings for the same word are stored. Of the shelf packages, if any, can be used.

---

<sup>2</sup>This will be elaborated more in the following components of the architecture.

In this process, the challenge of *alternative spellings* is addressed with respect to the *integrity of user's tags* and *integrity of social interaction patterns* criteria.

Having processed the tags, two copies of the user tags will be stored in the tagging system database. Further processing of the tags will be accomplished in the other components of our architecture.

So far, the first component of our architecture (tagging) addresses the following tagging system challenges:

1. Different lexical forms
2. Alternative spellings
3. Misspelling errors
4. Badly encoded tags

The processes accomplished in the *tagging component* complies with the *integrity of social interaction patterns* criterion. Furthermore, the changes in the tags do not affect the user-inserted tags. Rather, our architecture suggests replicating them in a refined manner. Therefore, it obeys the *integrity of user's tags* criterion. Other criteria are not relevant at this phase.

### 4.3.2 Searching component

*Searching component* is the highlighted area shown on the right hand side of our architecture in Figure 4.6.

Searching the tagging system is another form of user-system interaction. In fact, the processes suggested to be accomplished while searching vary depending on the type of

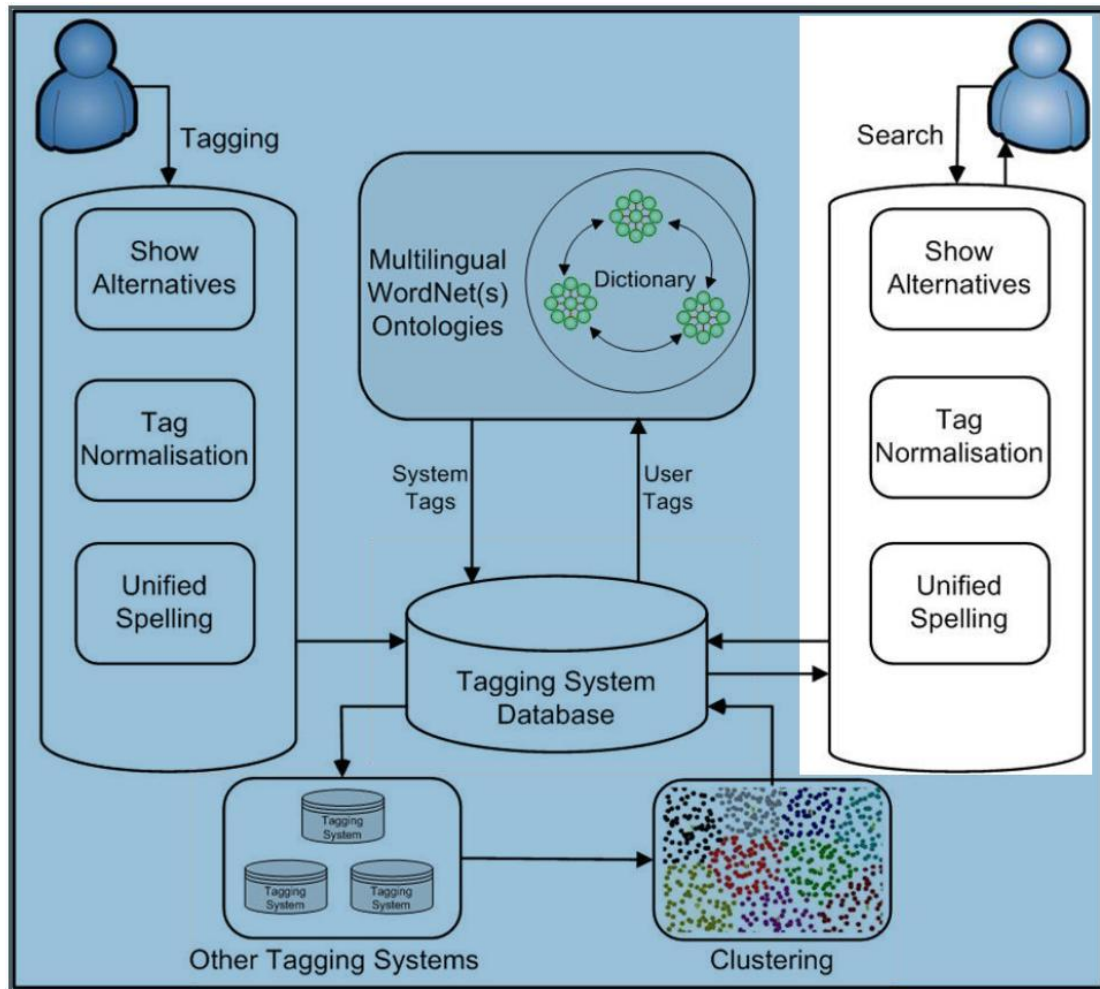


Figure 4.6: Generic architecture for tag-based systems - *Searching component*.

metadata that will be used in the search; *raw metadata* or *normalised metadata*. As aforesaid, the *raw metadata* are the original tags as inserted by the users without change, whilst the *normalised metadata* is the normalised copy of user tags.

### The first scenario

For search speediness reason<sup>3</sup>, the tagging system owner might decide to restrict the search to be in the *normalised metadata only*. In this case, the same processes accomplished in the *tagging component* should be accomplished in *searching component*; show-

<sup>3</sup>Rather comparing the efficiency of tagging systems in terms of time and space *with* and *without* system tags, here we only present some scenarios to minimise the use of time and space in case of adding system tags.

ing alternatives, normalising the user keywords, and unifying keywords spelling. The activity diagram of searching, using the *normalised metadata* only, is shown in Figure 4.7. The only difference between this activity diagram and the activity diagram for tagging (previously shown in Figure 4.5) is that the last activity in the tagging activity diagram is *storing in the database*, whilst it is *querying the database* in the searching activity diagram.

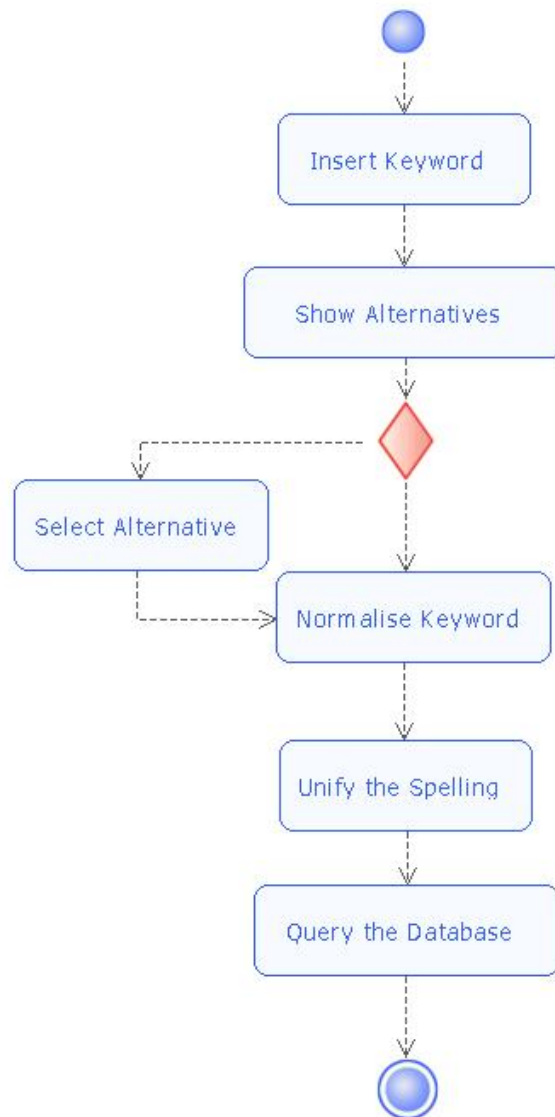


Figure 4.7: Searching activity diagram (using the *normalised metadata* only).

The reason behind this similarity is to guarantee that the normalised keywords used in

searching are matching the normalised keywords used in tagging. And so, some matching results can be retrieved. For example, if an original *user tag* was “*men*”, its normalised corresponding tag will be “*man*”. Future search trials will be in the *normalised metadata only*. Hence, the original tag “*men*” is not seen by the system’s search engine. Therefore, if another user is searching using the original keyword “*men*”, no result will be retrieved unless this keyword (“*men*”) is processed the same way it was processed at tagging time.

In this scenario, **only** the **normalised metadata** will be used. Consequently, the **normalisation** and the **spelling unification** processes are **necessary** for the searching keywords.

### **The second scenario**

On the other hand, if both *raw metadata* and *normalised metadata* are to be used, then the *normalisation* and the *spelling unification* processes at the searching time are not necessary. This is for the reason that future search trials will be in both metadata types. In this case, the original tag “*men*” (mentioned in the example above) is seen by the system’s search engine. Therefore, if another user is searching using the original keyword “*men*”, the tagged object will be retrieved. The activity diagram of searching, using both *raw metadata* and *normalised metadata*, is shown in Figure 4.8.

The decision to select either the *first scenario* or the *second scenario* depends on the hardware and software used in the system. In both cases there is a delay; in the first case, the delay will be in the processes of normalisation and spelling unification of the search keywords, but it will take less time while searching the database since it is only searching the normalised tags. In contrast, in the second case no time will be consumed for the processes of normalisation and spelling unification of the search keywords, but it will take more time while searching the database since it is searching both types of tags; the raw



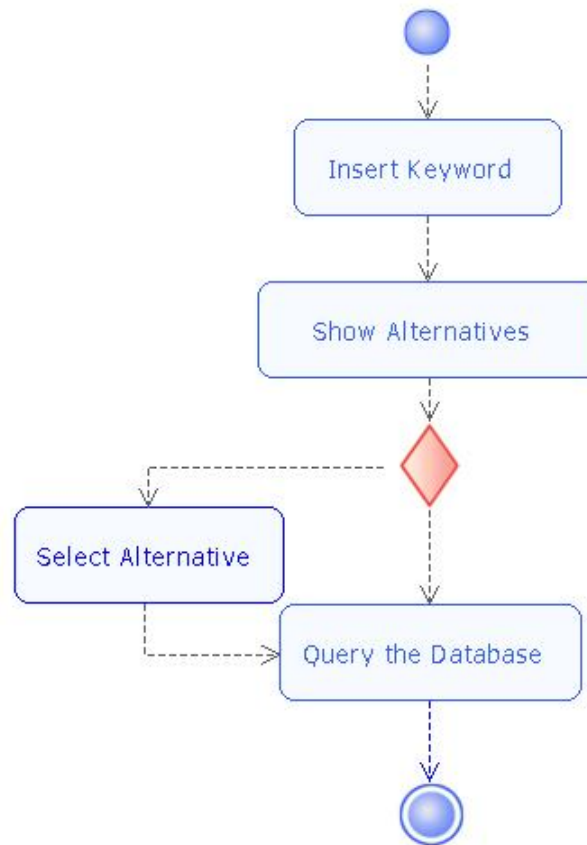


Figure 4.8: Searching activity diagram (using both *raw* and *normalised* metadata).

and the normalised tags.

In this scenario, **both** the **raw metadata** and the **normalised metadata** are used. Consequently, the **normalisation** and the **spelling unification** processes are **not necessary** for the searching keywords.

### The third scenario

In the previous two scenarios, we discussed the case when a user is searching the tagging system using a keyword that is identical with the original tag. And we explained that to make the original tag visible to the system's search engine, the choice is either to search in both metadata types without normalisation of the searching keyword, or to search only

in the normalised metadata with mandatory normalisation of the searching keyword.

In this scenario, we discuss the case when a user is searching the tagging system using a keyword that is not identical with the original tag, but it will be identical if it is normalised. In both examples used in the previous scenarios, the original *user tag* was “*men*” and the searching keyword was “*men*” also. What if the original *user tag* is “*man*” (the normalised version of the tag “*men*”) and the searching keyword is “*men*”? Table 4.2 shows the possible probabilities and the retrieved results for each of them, if any.

Scenario	Original tag	Normalised tag	Original searching keyword	Normalised keyword (if any)	Is there a result retrieved?
First Scenario	men	man	men	man	✓
	men	man	man	man	✓
	man	man	men	man	✓
	man	man	man	man	✓
Second Scenario	men	man	men	—	✓
	men	man	man	—	✓
	man	man	men	—	×
	man	man	man	—	✓

Table 4.2: Possible probabilities and the retrieved results for the *first* and the *second* scenarios.

We notice that the first scenario can retrieve results in all cases. Nevertheless, let us assume that there are two objects in the tagging system, one of them is originally tagged using the tag “*men*” whereas the other one is originally tagged using the tag “*man*”. In case of using the searching keyword “*men*”, the first scenario will retrieve the both objects without prioritisation. Similarly, the same non-prioritised results will be retrieved if the searching keyword is “*man*”.

In the second scenario, we note that if the original tag is “*man*” and the searching keyword is “*men*”, no results are retrieved since there is no normalisation for the searching

keywords.

In order to overcome the obstacles abovementioned for the first and the second scenario, we propose a third scenario. In this scenario, we suggest the search in both **raw metadata** and **normalised metadata** (similarly to the second scenario). On the other hand, we suggest the **normalisation** of the searching keywords (similarly to the first scenario). Although this scenario produces results in all the cases discussed before, with the higher priority results first, it needs more time for normalising the searching keywords and for producing results on both raw and normalised metadata.

In this scenario, the objects that were tagged with the exact searching keyword are expected to be more relevant results. Therefore, these objects will be retrieved first (prioritised results).

The differences, as well as the similarities, among the three proposed scenarios are briefly stated in Table 4.3.

Scenario	Both metadata are used	Searching keywords normalisation	Results prioritisation
First scenario	×	✓	×
Second scenario	✓	×	×
Third scenario	✓	✓	✓

Table 4.3: Comparison among the three scenarios.

In the *searching component* of our architecture, the stored tags are not affected by any kind of processing and no data is added to the tagging system database. Thus, none of the tagging system challenges were directly addressed. The processes accomplished for the searching keywords to match the normalised tags can be considered as a complementary

effort for addressing the same challenges that were addressed in the *tagging component*.

The relevant criterion for the *searching component* is the *Integrity of social interaction patterns* since there is user-system interaction in searching. This criterion is considered in this component of the architecture the same way as in the *tagging component*.

### 4.3.3 Semantic component

*Semantic component* is the highlighted area shown at the top-middle side of our architecture in Figure 4.9.

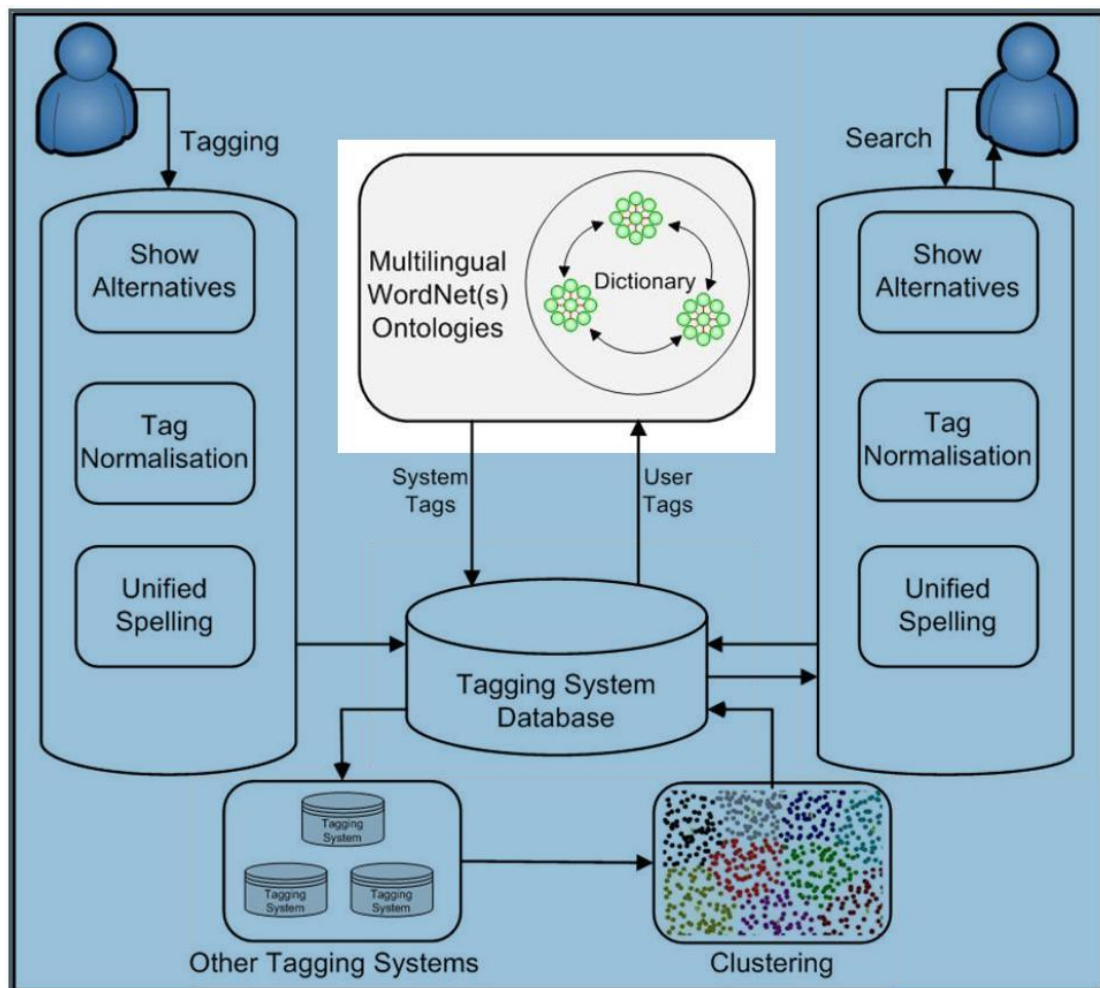


Figure 4.9: Generic architecture for tag-based systems - *Semantic component*.

As seen in Figure 4.9, there is no interaction with the user in the *semantic component*; the interaction is limited between the tagging system database and the embedded lexical semantic ontologies (WordNets). Thus, it is an internal interaction among the system's components.

Since we investigate tags only as a type of metadata in tagging systems, it is important to remind ourselves that our discussion from now on is assuming that tags are the only searchable metadata in tag-based systems.

### **Synonyms and semantic relations**

Lack of semantics among tags is one of the most prominent challenges in tagging systems. In this part of our architecture, we shed some light on the use of Semantic Web ontologies inside the tagging system. Thus, we make use of Web 3.0 technologies in Web 2.0 applications.

If a particular content in a particular tag-based system was tagged using the English word “*car*”, the unique keyword that can make access to that content in a future search is the word (“*car*”). In other words, if a user searches using the lexical synonyms of that word (e.g. “*auto*”, “*automobile*”, “*motorcar*”), that content will not be retrieved although it is relevant to the search terms. In a similar way, if a user searches using the lexical hypernyms, for example, of that word (e.g. “*motor vehicle*”, “*automotive vehicle*”), the same problem will arise; the content is related to the search terms but is not being retrieved.

As humans, we can understand the synonymy or hypernymy relation between these words, and therefore, take decisions according to this understanding. Tag-based systems cannot recognise such relations unless we make available some supplementary resources

to help clarify such relations.

The role of the semantic ontologies provided in our architecture is to add meanings to the *user tags* by consulting the WordNet ontologies. Based on the relations between words existing in WordNet, WordNet provides a set of words that are relevant to the *user tags*. This set of words will be saved in the tagging system database as *system tags*.

Once a new tag is inserted and stored in the tagging system database, the database will query WordNet using that tag. WordNet, in return, will give a set of *system tags* as a result. These *system tags* will be stored in the database as additional metadata describing the tagged content. The activity diagram explaining this process of interaction between the tagging system database and the semantic ontologies (WordNets) is shown in Figure 4.10.

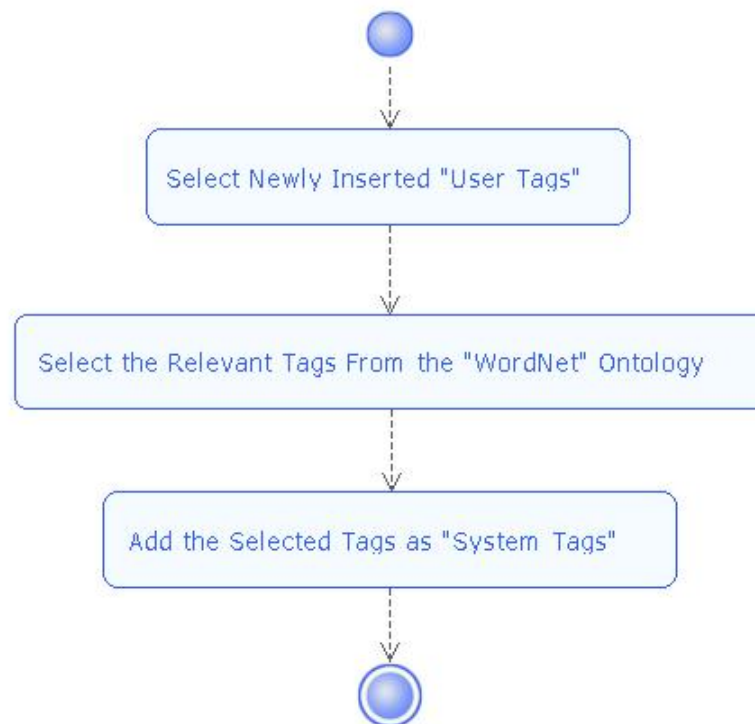


Figure 4.10: The activity diagram of adding *system tags* from the semantic ontologies.

Consequently, in the abovementioned example, the tagged object will not only be accessible by the original *user tag* (“*car*”), but also by all other relevant *system tags* (“*auto*”, “*automobile*”, “*motorcar*”, “*motor vehicle*”, and “*automotive vehicle*”).

Relevant words that can be used as *system tags* might be the word *synonyms* and *hypernyms*. However, more investigations and experiences are needed to agree which types of relevant words are the most appropriate choices. Here we mention some possible scenarios to select the relevant tags from the WordNet(s), and thus, to add as *system tags*:

- Adding only the synonyms of the *user tag* as *system tags*.
- Adding only the synonyms and the direct hypernyms (one level of hypernyms) of the *user tag* as *system tags*.
- Adding the synonyms and two levels of hypernyms (the hypernym of the hypernym) of the *user tag* as *system tags*. Likewise, three or more levels might be suggested.
- Adding the synonyms, hypernyms, and the synonyms of the hypernyms of the *user tag* as *system tags*.

Anyhow, such scenarios need to be examined in real systems and their results need to be analysed and evaluated to make the right decision (we will do part of this in the following chapters).

### **Multilinguality**

Users of tag-based systems use different languages for tagging and searching. Furthermore, taggers in tagging systems are not necessarily using the same language as searchers. Nevertheless, the content that was tagged using a specific language is not accessible unless identical search terms of the same language are used; the tagging systems cannot translate the search terms. For example, to find a content that was tagged using English language,

searchers should use matching English search terms; corresponding search terms of any other language retrieve naught.

In the abovementioned example of “*car*”, the tagged object is accessible only by using the English relevant words of the word “*car*” (after applying the suggestions in our architecture). If an Italian user, for example, searches using any of the words “*autovettura*”, “*vettura*”, or “*macchina*”; which are the Italian equivalent of the English word “*car*”, no results will be retrieved. This is a real challenge; the relevant content is there but not accessible due to translational barriers.

Back to Figure 4.9, we notice that the *semantic component* contains *multilingual semantic resources*. The *multilinguality* dimension in our architecture adds translational functionality to the *user tags* in form of *system tags*. Based on the WordNet structure and relations, a number of multilingual lexical ontologies are available nowadays such as MultiWordNet and EuroWordNet ontologies. These ontologies contain a number of languages saved in one place (usually a database) with a cross-language linkage among the word translations in different languages. Therefore, bringing such multilingual lexical resources in the tagging system can help addressing the challenge that taggers and searchers are using multiple languages.

Akin to querying the WordNet previously shown in Figure 4.10, the multilingual lexical ontology can be queried using the newly inserted *user tag*. In this case, the *system tags* from the multilingual ontology will contain relevant words in different languages.

Consequently, in the abovementioned example, in case that the Italian language is included in the multilingual ontology, the tagged object will be accessible using the Italian system tags “*autovettura*”, “*vettura*”, and “*macchina*”.



Nowadays, there is a WordNet version available for many languages. Although some of these WordNets are grouped together in one place as a multilingual resource (e.g. MultiWordNet, EuroWordNet), other WordNets are still available independently. The translation among the languages that are existing in one multilingual resource is accomplished via the cross-language linkage. However, such linkage does not exist among the independent WordNets which makes the translation among these WordNets more complex. Therefore, some online dictionaries can be used as intermediary among these individual WordNets.

The tagging system needs to make automatic use of the online dictionaries. The challenge in this case is the determination of the source language of a given user tag. Usually, taggers are registered users; their profile information are available for the tagging system. Therefore, some profile information (e.g. nationality, country, etc) might give a clue in determining the language used in the user tags. Furthermore, some extra information can be obtained by the users at registration time for the purpose of determining the language(s) that will be used in tagging.

### **Alternative Scenario**

The scenario of adding *system tags* in the *semantic component* is arguable. Our suggested scenario is to accomplish the computational processes (e.g. consulting the ontologies, searching for the relevant words, etc) at the *tagging time*; once the tags are inserted by the user. Therefore, no computations, which are ***time consuming***, are needed at the *searching time*. Given that each *user tag* has more than one *system tag* in most cases, adding *system tags* is a ***space consuming*** choice.

In contrast, another scenario can suggest abandoning the *system tags* to save the space and accomplishing all the computations at the *searching time*. The activity diagram shown

in Figure 4.11 explains the alternative scenario that can occur at *searching time*.

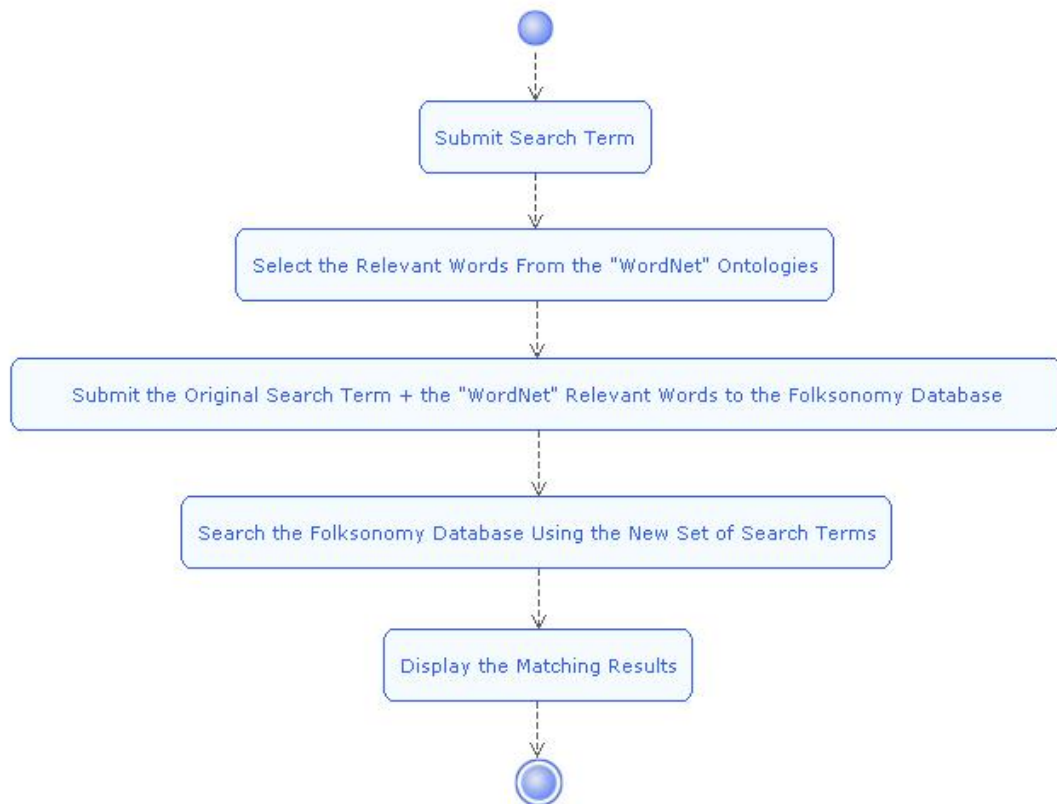


Figure 4.11: Alternative scenario activity diagram.

In the abovementioned example of “*car*”, the content was tagged using the English word “*car*”. The alternative scenario suggests no action to be taken at *tagging time*. Alternatively, when another user submits the search term “*automobile*”, the WordNet will be queried *first* to retrieve all the relevant words of the given search term “*automobile*”. The result set of querying the WordNet will contain the words “*car*”, “*auto*”, “*motorcar*”, “*motor vehicle*”, and “*automotive vehicle*”<sup>4</sup>. The results of querying the WordNet, as well as the original search term, will be used as new search terms to search the tagging system database. Therefore, any content that was originally tagged using any of these search terms will be retrieved. Consequently, the object that was tagged as “*car*” will be

---

<sup>4</sup>Assuming that the results will include only the English synonyms and the direct hypernyms.

retrieved. This scenario should happen behind the scene so that the user will not notice any change on the submitted search terms.

The critical factors to consider in deciding which scenario to adopt are *time* and *space*. We argue that storing *system tags* (the first scenario) is more efficient than the alternative scenario for the following reasons:

1. General speaking, *time* is the more significant factor that comes in the first place in searching, *space* comes in the second place. The first scenario saves more *time* but consumes more *space*.
2. Computations, which are time consuming, can occur either at tagging time (the first scenario) or at searching time (the alternative scenario). *Time* is more important at searching time because users are waiting for a response, while it is less significant at tagging time. Furthermore, computations will take place when the user finishes the tagging process and leaves the system; no response is expected at tagging time.
3. The *system tags* are not consuming much extra space due to their nature; tags are textual data which occupy trivial amount of space in the tagging system database.

The *semantic component* of our architecture addresses the following tagging system challenges:

1. Word synonyms
2. Semantic relations
3. Multilinguality

The relevant criteria that were considered throughout the *semantic component* processes are listed below:

- *Integrity of user's tags*: The added *system tags* are not replacing or changing the *user tags*. Furthermore, taggers will not see the added *system tags* with their tags; *system tags* are visible only for the internal components of the tagging system for searching purposes.
- *Integrity of social interaction patterns*: Even though the search terms were processed in the alternative scenario (if adopted), the user-system interaction will not be changed at searching time.
- *Rich functionality*: The suggested *semantic component* can be used with existing systems; we are not proposing an entirely new tagging system. The functionalities added in this component are accomplished in the background of the system to enhance the search accuracy.
- *Multilinguality*: By using different WordNet ontologies, the tagging system can recognise different languages. Therefore, the content of the tagging system is accessible for users regardless of their languages.

#### 4.3.4 Clustering component

*Clustering component* is the highlighted area shown at the bottom-middle side of our architecture in Figure 4.12.

Similarly to the *semantic component*, the *clustering component* has no interaction with the user; the interaction is limited between the tagging system database and other tagging systems' databases (at least one other tagging system database). These databases should be kept as external social corpuses; they should not be integrated as part of the current tagging system. Whenever needed, they will be contacted.

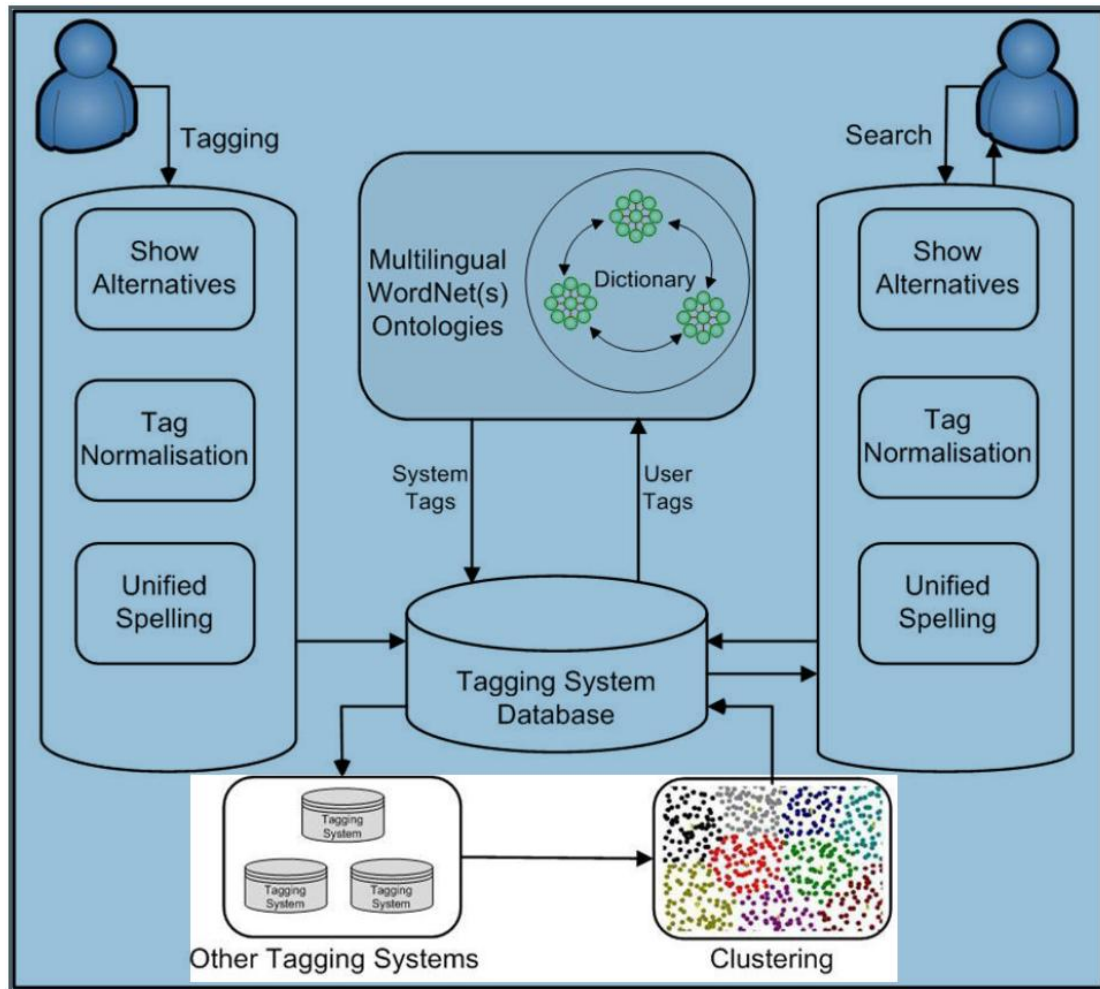


Figure 4.12: Generic architecture for tag-based systems - *Clustering component*.

One of the drawbacks of using WordNet is the existence of tags that are not recognised in the WordNet’s lexicon [6]. In tagging systems, users tend to use special language in tagging; shorthands, colloquial words, or specialised technical terms in different domains of knowledge (e.g. “XML”, “RSS”, “folksonomy”). Such tags will not be found in the *semantic component* of our architecture. Therefore, another resource should be found to add semantics to these tags.

The emergent tags that cannot be found in the lexical ontologies are usually existing in the Social Web; ontologies do not recognise the meaning of the shorthand “luv” whilst

it has a meaning in the online social community. The rationale behind this statement is the fact that such tags have emerged at first in the Social Web community. Therefore, their meanings can be extracted only from the Social Web community itself. Analysing the use of such tags in the tag-based systems can give a clue about their semantics. We suggest *clustering* for analysing such tags in the tag-based systems.

In clustering, the dataset of tags in a particular tagging system can be grouped in smaller subsets of tags called *clusters*. A group of tags can be put together in one cluster based on predefined criteria (e.g. tags co-occurrence). The existence of a group of tags in the same cluster implies a semantic relation among these tags from the users' viewpoint. This relation can be used to add semantics to *user tags* that were not found in the semantic ontologies.

Adding semantics to the *user tags* that were not found in the *semantic component* takes place in two phases. The first phase is to contact the external tagging system(s), perform the clustering process, and save the resulted tag clusters in the *database component* of our architecture (see the activity diagram in Figure 4.13). This can be accomplished periodically to capture the latest unsupervised social vocabulary and to keep the tag clusters up-to-date.

Having finished this phase, each tag in the clustered tag-based system belongs to one cluster saved in the *database component*.

All the tags obtained from the external tagging system(s) have meanings since they are semantically correlated to other tags inside the same cluster. Grouping tags in clusters gives a *context* for each tag that might help in defining the ambiguous tags.

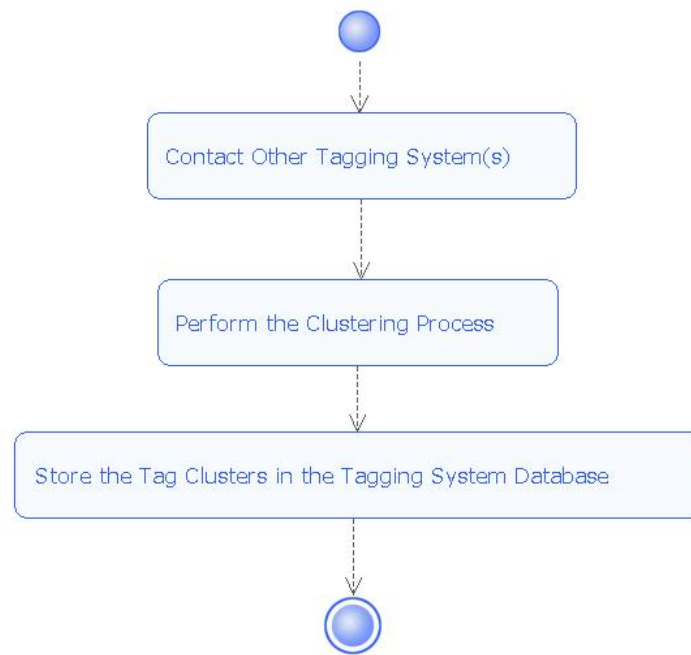


Figure 4.13: Clustering activity diagram.

The second phase is the use of the stored tag clusters, this will be accomplished once a new *user tag* is saved in the tagging system database. In case that the inserted *user tag* has no meaning in the *semantic component*, the tagging system will, alternatively, search for that tag in the clusters obtained in the previous phase. Once a cluster for the new tag is found, the *Top-N* related tags in that cluster will be picked and saved as corresponding *system tags* for the tagged object. The activity diagram of adding *system tags* using the tag clusters originally obtained from external tagging system(s) is shown in Figure 4.14.

The tag clusters that are stored in the current tagging system database can be filtered for better efficiency in terms of space and time. These clusters are used *only* to find the meanings of the tags that do not have meaning in the WordNet(s). In other words, if a tag has a meaning in the WordNet(s), its *system tags* will be added from the *semantic component* and no *system tags* will be added from the stored clusters. Therefore, if a tag has a meaning in the *semantic component*, its cluster will never be used and thus can be

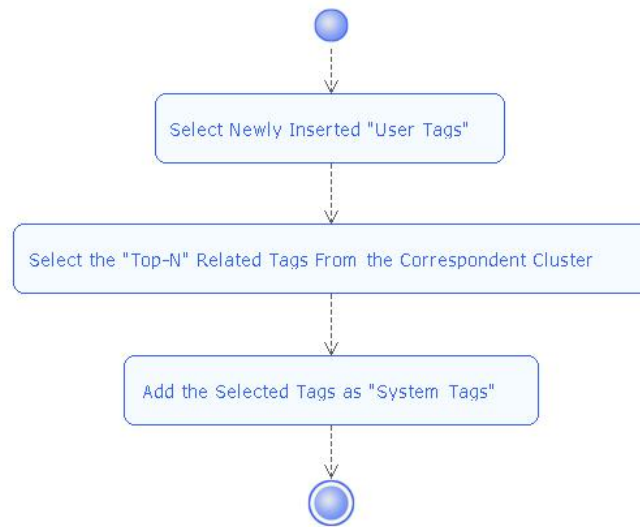


Figure 4.14: The activity diagram of adding *system tags* from the tag clusters.

deleted from the current tagging system database.

The activity diagram shown in Figure 4.15 describes the process of adding *system tags* either from the *semantic component* or from the tag clusters obtained from the *clustering component* (The first diagonal box tests if the tag exists in the WordNet(s), whilst the second one tests if the tag has a corresponding cluster).

### Alternative scenario

To our knowledge, the Flickr tagging system provides *tag clusters* which contain a set of Flickr users' tags put together in groups. The criteria used to group these tags together have not been released officially by Flickr [110]. In [8], they claim that Flickr has grouped the tags in clusters according to the relatedness of tags, although these tags fall into several semantic categories. Others claim that the clusters group similar terms together [111]. Our architecture can benefit from available tag clusters to save time and space.

Flickr provides various APIs that enable other applications to make use of its clusters. Given a particular tag, Flickr will give the related cluster(s) for that tag. Here we list the



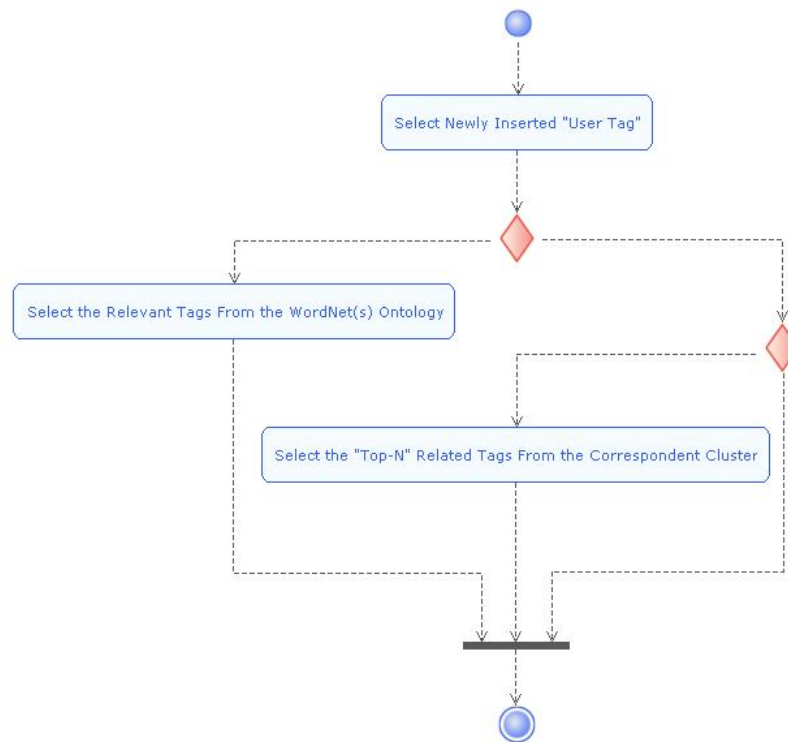


Figure 4.15: The activity diagram of adding *system tags*.

clusters retrieved from Flickr for some tags that have no meaning in the WordNet. The number of tags varies from one cluster to another. In our examples, we pick the *Top-10* tags in each cluster (if there are more than 10 tags in the cluster).

- “*luv*” has three clusters:

Cluster 1 = {love, heart, red, amor, pink, canon, girl, green, happy, rose}

Cluster 2 = {boeing, southwestairlines, airplane, southwest, swa, airport, jet, boeing737, aviation}

Cluster 3 = {bjd, doll, dollmore}

- “*txt*” has one cluster:

Cluster 1 = {phone, text, sms, mobile, message, cellphone, cell, texting, textmessage, msg}

- “*folksonomy*” has one cluster:

Cluster 1 = {tagging, tags, web20, flickr, delicious, tag, technorati, tagcloud, extispicious, ajax}

- “XML” has four clusters:

Cluster 1 = {rss, atom, macintosh, apple, mac, feed, feeds, itunes, cms}

Cluster 2 = {books, oreilly}

Cluster 3 = {css, xhtml, design, web, flash, webdesign, html, portfolio, blog, actionscript}

Cluster 4 = {ajax, java, screenshot, web20, javascript, flickr, linux, linux, mysql, programming}

Having a look at the clusters in the examples above, we can notice the following:

1. Different number of clusters is retrieved for each tag (e.g. three clusters for the tag “luv”, and one cluster for the tag “txt”).
2. Each cluster presents different sense of the tag (e.g. “luv” in *Cluster 1* means “love”, while it refers to “an airlines company” in *Cluster 2*).
3. Within a particular cluster, some tags (usually the *Top-N* tags) are related tags (e.g. *Cluster 1* for the tag “luv”, *Cluster 3* for the tag “XML”).
4. The order of the related cluster is not known; sometimes it is the first cluster, but not necessarily (e.g. *Cluster 3* for the tag “XML”).

So far, the relatedness of the clusters to a given tag can be determined manually. Revealing the criteria used by Flickr can help choosing the best cluster for a given tag in an automatic way. We argue that the related clusters give a robust context for a given tag. Therefore, the *Top-N* tags in the most related cluster can be used as *system tags* for the given tag. These tags might help in clarifying the meaning and add semantics to the ambiguous tags that do not exist in the WordNet ontologies.

The *clustering component* addresses the challenge of *shorthand writing*. In addition to the *Integrity of user's tags*, *Integrity of social interaction patterns*, and *Rich functionality* criteria, the most relevant criterion in this component is the *dynamism* since the clusters are being updated periodically.

### 4.3.5 Database component

*Database component* is the highlighted area shown in the middle of our architecture in Figure 4.16.

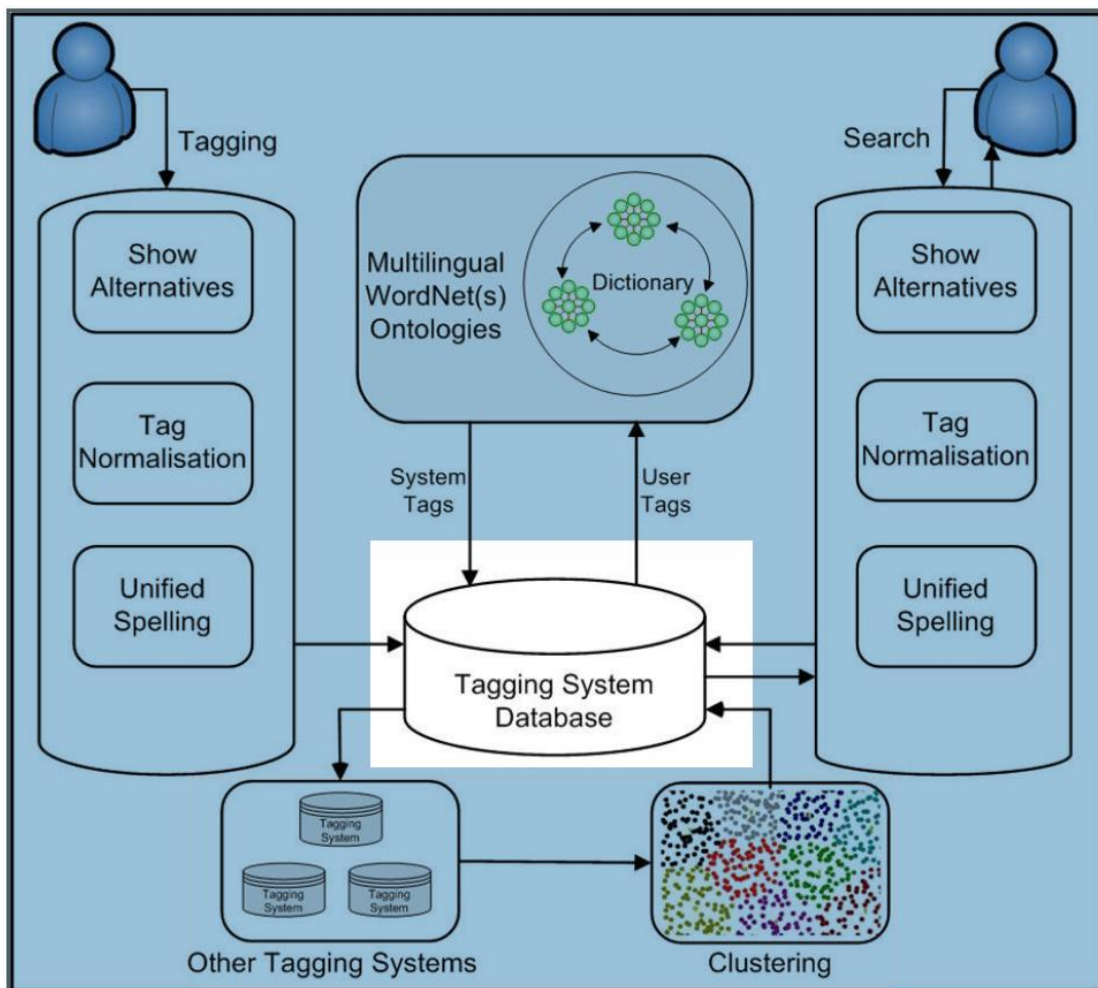


Figure 4.16: Generic architecture for tag-based systems - *Database component*.

The tagging system database is the place where all the information about user gener-

ated content is stored, including the data and the metadata. The design of the database might vary from one tagging system to another depending on the nature and specificity of each tagging system. What our architecture suggests is a general design for the tagging system database regarding the tags storage; which represent the central part of the meta-data in tag-based systems.

The suggested design of the tagging system database should be tuned with the requirements aforesaid about the tags replication in one hand, and on the other hand, it should conform with the other components of the architecture. As we see in Figure 4.16, the *database component* has contact with all the other components of the architecture; the *tagging component* delivers *raw* and *normalised* tags to be stored in the database, the *searching component* queries and retrieve results from the database, and the other two components store *system tags* in the database.

The *database component* might contain many tables to store the data of users, tagged objects, clusters, etc. Our concern here is the tables where *user tags* and *system tags* will be stored. A proposed outline for these tables structure is shown in Figure 4.17.

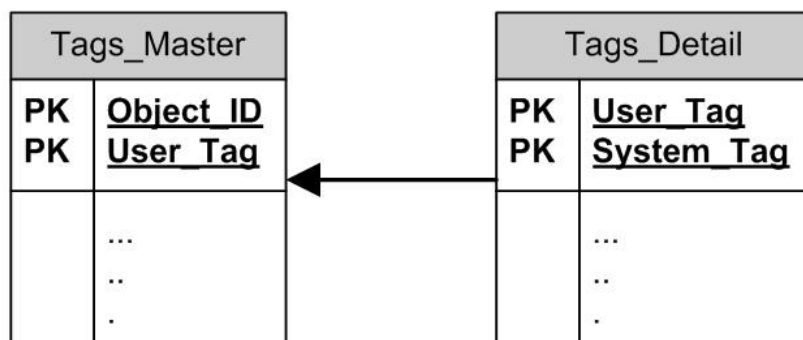


Figure 4.17: Logical diagram for the *tags* tables.

This structure guarantee the minimal redundancy of data; the *system tags* for a specific *user tag* are uniquely stored in the database. In other words, one copy of the *system tags*

for the tag “*beautiful*”, for example, will be stored in the table “Tags\_Detail” once a user inserted that tag for the first time. Future use of the tag “*beautiful*” by other users will not replicate its *system tags*.

## 4.4 Summary of our architecture

The architecture has promising features to address the current challenges of tagging systems. Yet, it has some limitations.

### 4.4.1 Virtues

The criteria mentioned earlier in this chapter were concluded from a deep investigation in the literature of tag-based systems. These criteria are respected in the architecture as aforesaid for each component.

Furthermore, the architecture addresses most of the aforementioned tagging system challenges. Table 4.4 summarises these challenges; the first column designates the challenge whilst the other one specifies the component of our architecture which addressed the challenge, if addressed<sup>5</sup>.

### 4.4.2 Limitations

Few challenges come into view in addressing *multilinguality* and *shorthand writing*; namely, in the *semantic* and *clustering* components respectively. Specifying the source language of *user tags* is a prerequisite for translating these tags into different languages. We suggested adding more information to the users’ profiles to help identifying the language(s) they might use in their tags. Yet, more investigation is needed.

---

<sup>5</sup>The unaddressed challenges are beyond the scope of our work.

The challenge	The component	
Word synonyms	✓	Semantic component
Word polysemy	×	Not addressed
Different lexical forms	✓	Tagging component
Alternative spellings	✓	Tagging component
Misspelling errors	✓	Tagging component
Badly encoded tags	✓	Tagging component
Specialised tags	×	Not addressed
Key phrases instead of keywords	×	Not addressed
Multilinguality	✓	Semantic component
Shorthands	✓	Clustering component
Semantic relations	✓	Semantic component

Table 4.4: The addressed challenges in our architecture.

Clustering other tagging system(s) needs particular permissions and agreements between the *clustering tagging system* and the *clustered tagging system(s)*. This is a business-related issue since the authorities behind these tagging systems are, most likely, independent organisations.

## 4.5 Summary

Having investigated the advantages and challenges of tagging systems, we formulated criteria for a robust tagging system. These criteria consider the integrity of user tags, the integrity of social interaction pattern, the rich functionality, the universality, the dynamism, and the multilinguality.

With respect to these criteria, we built a generic architecture for tag-based systems consisting of five components; tagging, searching, semantic, clustering, and database component. Most of the known challenges of tagging systems, as well as the shorthand writing challenge and the semantic relations challenge that we introduced, are addressed

by the proposed architecture. The architecture suggests automatic adding of tags by the tagging system using some internal and external resources, so-called *system tags*, to disambiguate the *user tags*.

## **Part III**

# **Implementation and Evaluation**



## Chapter 5

# Prototype Implementation of the Tagging System Architecture

### *Objectives:*

---

- Developing a prototype implementation of the semantic and clustering components of the tagging system architecture.
  - Discussing the rationale for using the semantic resources and the social resources in the prototype.
  - Describing the algorithm of adding system tags in tag-based systems.
-

## 5.1 Introduction

In the previous chapter, we introduced a generic architecture for tagging systems which addresses most of the tagging systems challenges. A prototype implementation is presented in this chapter with a narrower scope than the scope previously discussed in the architecture. Therefore, the challenges that will be addressed in the prototype implementation is a subset of the challenges that were addressed in the architecture.

In the prototype implementation, we will zoom-in on two components of the generic architecture; the *semantic component* and the *clustering component*. Therefore, more elaboration about the sub-components of these two components will be presented in this chapter.

Further discussion about the implementation design will be provided including the database design and the interaction between the database and the other two implemented components; the semantic and the clustering components. The algorithm we developed and used for adding system tags in tag-based systems will be presented in a pseudo code form.

## 5.2 The prototype scope

As aforesaid, the main aspects of improvements in our architecture are the semantic aspect, the multilinguality aspect, and the clustering aspects. These aspects lie in two components of the architecture; the *semantic component* and the *clustering component*. Therefore, the prototype will implement these two components. Figure 5.1 shows the tagging system architecture with numbered labels, the prototype implementation will deal with the components 3, 4, and 5.

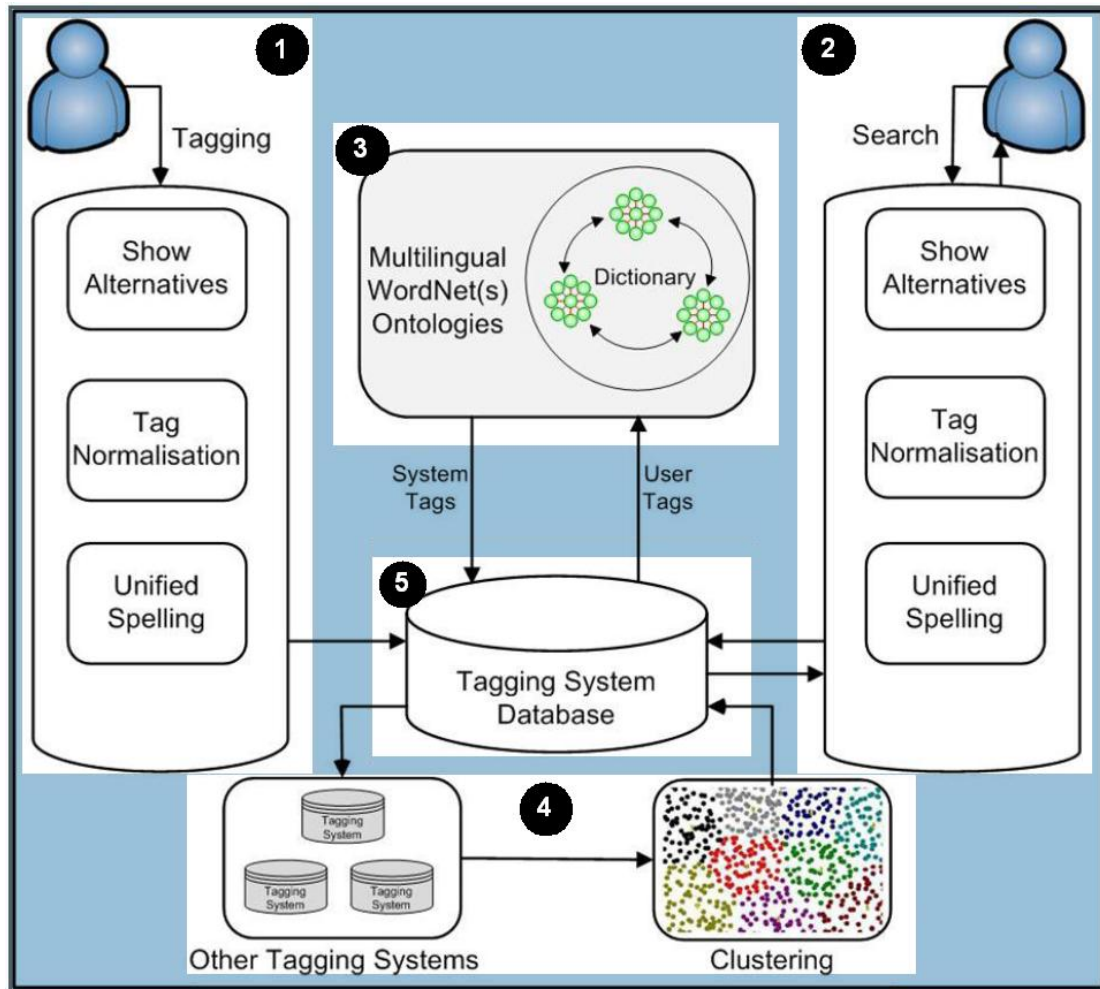


Figure 5.1: The components of the tagging systems architecture.

In other words, the prototype is focusing on adding system tags from two different resources; the semantic resources (e.g. WordNet ontology) and the clusters obtained from other tagging systems. Hence, the system tags will be acquired from two web generations; the Semantic Web and the Social Web.

### 5.2.1 Semantic resources

As discussed in the previous chapter, the semantic component of the generic architecture for tagging systems addresses the challenges of *word synonyms*, *semantic relations*, and *multilinguality*. In order to address the challenges of *word synonyms* and *semantic*

*relations*, the PWN ontology is used. For the *multilinguality* challenge, there are different multilingual semantic ontologies such as MWN, and EuroWordNet. The former contains two languages; English and Italian, whereas the latter contains seven European languages; Dutch, Italian, Spanish, German, French, Czech and Estonian. In this work, we used MWN as it is available free for researchers. Figure 5.2 shows the semantic component in our prototype.

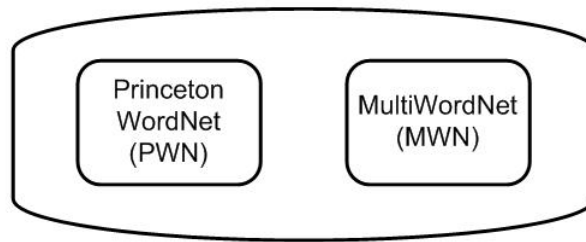


Figure 5.2: The semantic component in our prototype.

As aforementioned in the previous chapters, the PWN and MWN ontologies are similar in the structure and have the same kinds of relations among words. In our work, these relations are used to determine which words in the ontology (PWN or MWN) are relevant to the user tag, and hence, to be added to the tagging system database as *system tags*.

Here we discuss the main relations, and some other important subordinate relations, in PWN and MWN and demonstrate the rationale behind our decision of which relations can be included, or excluded, for adding system tags. The main relations are *synonymy*, *antonymy*, *hyponymy/hypernymy*, *meronymy*, *entailment*, and *troponymy*<sup>1</sup>. The subordinate relations are “*similar to*” and “*also-see*” relations.

It is significant to remind here that in PWN, and thus MWN, each word has many *senses*; which are different meanings for the same word. *Senses* in PWN, and other ontologies which are based on PWN structure, are generally ordered from most to least

---

<sup>1</sup>As aforementioned in Figure 2.1.

frequently used, with the most common sense listed first and so forth [112]. The same word might have different kinds of relations based on the *sense* to which it belongs. For example, the verb “run” has 41 *senses* in PWN. “Run” in the *first sense* means “move fast by using one’s feet, with one foot off the ground at any given time” and has no synonyms, while in the *seventh sense* it means “perform as expected when applied” and has four synonyms (“function”, “work”, “operate”, and “go”). In the following relations, we always deal only with the *first sense* which is the most frequent one.

1. Synonymy: By definition, two words are considered to be synonyms if the substitution of one for the other in a linguistic context will seldom alter the meaning in that context [22]. The definition indicates that the synonyms for a specific word give equivalent meanings and can be used interchangeably to refer to the same meaning. For example, “child”, “kid”, “youngster”, and “nipper” are used interchangeably to refer to a young person of either sex. If a user tagged an object using the word “child”, it is sensible to add the other synonyms by the system as “system tags” since it is reasonable for a searcher that is using the word “kid” to retrieve objects that were originally described (tagged) by the tagger using the word “child”. Therefore, we decided to **include** this relation in adding the system tags.
2. Antonymy: There is no similarity between the meaning of a word and its antonym. Rather, the antonym of a word gives the opposite meaning of that word’s meaning; a word and its antonym are not interchangeably used to refer to the same meaning. For example, the antonym of “clean” is “dirty”. If a user tagged an object using the word “clean”, it is not sensible to add its antonym “dirty” by the system as “system tag” since it is not reasonable for a searcher that is using the word “dirty” to retrieve objects that were originally tagged using the word “clean”. Therefore, we decided to **exclude** this relation in adding the system tags.
3. Hyponymy/hypernymy: Hyponymy and hypernymy are two different readings for

the same relation. That is; if  $x$  is a hyponym of  $y$ , then  $y$  is a hypernym of  $x$ . For example, “*love*” is a hyponym of “*emotion*”, and thus, “*emotion*” is the hypernym of “*love*”. In other words, “*love*” is kind of “*emotion*” and there might be other kinds of “*emotions*” such as “*hate*”, “*anger*”, etc. To decide whether to include or exclude this relation in adding system tags, we should think of the two readings for this relation; namely, hyponymy and hypernymy.

If a user tagged an object using the word “*love*”, it is sensible to add the hypernym “*emotion*” by the system as “*system tag*” since it is reasonable for a searcher that is using the word “*emotion*” to retrieve objects that were originally tagged using the word “*love*”. Indeed, “*love*” is a specialisation of “*emotion*”, and vice versa, “*emotion*” is a generalisation of “*love*”. When a general search term is submitted in a search, it is accepted (and might be expected) to retrieve results that were tagged using specialised terms, or kinds, of that general term. Therefore, we decided to **include** hypernymy relation in adding the system tags.

For the opposite reading (hyponymy), when a specialised search term is submitted in a search, it is not accepted (nor expected) to retrieve results that were tagged using generalised terms of that specialised term. For example, if a user tagged an object using the word “*emotion*”, it is not sensible to add the hyponyms (“*love*”, “*hate*”, “*joy*”, “*anger*”, etc) by the system as “*system tags*”. This is due to the fact that it is not reasonable for a searcher that is using the word “*love*” or “*hate*” (which are antonyms) to retrieve the same objects that were originally tagged using the word “*emotion*”. Since “*love*” and “*hate*” have opposite meanings, they should retrieve different result sets. Therefore, we decided to **exclude** hyponymy relation in adding the system tags.

4. Meronymy: The meronymy of the noun  $x$  is the noun  $y$  where  $y$  is part of  $x$ . For example, “*hand*”, “*arm*”, and “*face*” are some meronyms (parts) of “*man*”. If a user tagged an object using the word “*man*”, it is not sensible to add its meronyms (“*hand*”, “*arm*”, “*face*”, etc) by the system as “*system tags*” since it is not reasonable for a searcher that is using the word “*hand*” to retrieve objects that were originally tagged using the word “*man*”. Therefore, we decided to **exclude** meronymy relation in adding the system tags.
5. Entailment: This relation is only between verbs. For example, “*walk*” entails “*step*”, but not the other way around. Among 12,144 verbs in the MWN, only 429 verbs have this relation (only 3.5% of the verbs). If a user tagged an object using the word “*walk*”, it is sensible to add its entailed verb “*step*” to the system as “*system tag*” since it is reasonable for a searcher that is using the word “*step*” to retrieve objects that were originally tagged using the word “*walk*”. But on the other hand, this is not the case all the time; most of the time the entailment relation is misleading. According to our observation, the verbs that have this relation rarely have it for the *first sense* which is the most frequently used. Rather, in most of the cases the entailment relation exists for the less frequently used senses. For example, the verb “*go*” does not entail any other verb in its *first fourteen senses*, and it entails “*be*” in its *fifteenth sense* which means “*continue to live, endure or last*”. Also, the verb “*carry*” does not entail any other verb in its *first thirty-nine senses*, and it entails “*conceive*” in its *fortieth sense* which means “*be pregnant with*”. Since we restricted our work to deal only with the first sense, we decided to **exclude** entailment relation in adding the system tags. Nevertheless, it can be examined and evaluated in other empirical experiments<sup>2</sup>.
6. Troponymy: Similar to hyponymy relation between nouns, troponymy relation is

---

<sup>2</sup>The examples are taken from PWN.

between verbs. Hyponymy is read as “*is-kind-of*” whilst troponymy is read as “*is-manner-of*”. Therefore, we decided to **exclude** this relation in adding the system tags for the same reasons mentioned above in the hyponymy relation.

For the other two subordinate relations:

1. “Similar to”: As explained before, the non-antonymous adjectives are grouped in clusters around the antonymous adjectives. Each cluster contains adjectives that are similar in meanings but not close enough to put in one synset as synonyms (according to WordNet’s rules). For example, the following groups of adjectives have similar meanings but not similar enough to be synonyms:

- (“*wet*”, “*watery*”, “*moist*”, “*damp*”, “*humid*”, “*soggy*”, “*bedewed*”)
- (“*dry*”, “*parched*”, “*arid*”, “*dried*”, “*sere*”, “*withered*”, “*rainless*”)
- (“*beautiful*”, “*beauteous*”, “*gorgeous*”, “*pretty*”, “*splendid*”, “*glorious*”)
- (“*aggressive*”, “*assertive*”, “*hostile*”, “*truculent*”, “*bellicose*”, “*combative*”)
- (“*ambitious*”, “*aspirant*”, “*manque*”, “*wishful*”)

Classification of words in PWN is well accurate, and has extremely strict rules for deciding whether two adjectives are *synonyms* or just *similar to* each other. In tag-based systems, we argue that the selection of appropriate tags by the users is less accurate; there is no big difference between the words “*gorgeous*” and “*glorious*” while tagging a picture of a “*beautiful*” girl<sup>3</sup>. The same slighness of difference between similar words takes place at the search time. In other words, if a user tagged an object using the word “*beautiful*”, it is sensible to add its similar adjectives (“*gorgeous*”, “*glorious*”, “*pretty*”, “*splendid*”, etc) by the system as “*system*

---

<sup>3</sup>In the well-known and most popular text editor Microsoft Word, a group of words that are not considered to be synonyms in PWN, are considered to be synonyms. The synonyms which are suggested by Microsoft Word are widely accepted and used by users.



*tags*” since it is reasonable for a searcher that is using the word “*pretty*”, for example, to retrieve objects that were originally tagged using the word “*beautiful*”. Therefore, we decided to **include** “*similar to*” relation in adding the system tags.

2. “Also-see”: This relation is also called *related terms*. Although there are no revealed criteria (to our knowledge) for judging whether two adjectives are *related* to each other or not, this relation shows high relatedness in meaning between the adjectives linked by this relation. Anyhow, there is an intersection between the “*also-see*” relation and the “*similar to*” relation; some adjectives are linked using both relations at the same time as will be shown in the examples. The following groups of adjectives are linked by the “*also-see*” relation:

- (“*aggressive*”, “*assertive*”, “*hostile*”, “*offensive*”)
- (“*beautiful*”, “*attractive*”, “*graceful*”, “*pleasing*”)
- (“*happy*”, “*cheerful*”, “*glad*”, “*joyful*”, “*joyous*”)
- (“*dangerous*”, “*insecure*”, “*vulnerable*”)
- (“*nice*”, “*pleasant*”)

Yet again, we notice the slight difference among the meanings of the adjectives in each group. In tagging systems, if a user tagged an object using the word “*happy*”, it is sensible to add its related adjectives (“*cheerful*”, “*glad*”, “*joyful*”, “*joyous*”) by the system as “*system tags*” since it is reasonable for a searcher who is using the word “*cheerful*”, for example, to retrieve objects that were originally tagged using the word “*happy*”. Therefore, we decided to **include** “*also-see*” relation in adding the system tags.

As far as this, we discussed which relations of the PWN and MWN will be included (or excluded) in adding the system tags. The relations to be included are:

1. *Synonymy* relation (synonyms).

2. *Hypernymy* relation (hypernyms).
3. *Similar to* relation (similar terms).
4. *Also-see* relation (related terms).

### **Transitivity of relations**

Transitivity of relations refers to the levels of inclusion to be considered for the above specified relations. If a *user tag* is inserted, particular “rules” should be defined to limit the using of the specified relations in adding the *system tags*. The “rules” should answer the following questions:

- How many levels of hypernyms will be considered for adding the system tags; the first hypernym should be added only, or this might extend to the second, or even higher, level? For example, will the hypernym of the hypernym of a *user tag* be added also as a *system tag*?
- Will the synonyms, similar terms, and related terms be applied only for the user tag, or it will be also applied for the hypernym(s) of the user tag? In other words, will the synonyms, for example, of a given *user tag* be added only as *system tags*, or also the synonyms of the hypernym of that *user tag* will be added as *system tags*?
- Will the similar and related terms of a given *user tag* be added only as *system tags*, or also the hypernyms of these similar and related terms will be added as *system tags*?

There are no standards suggested for such rules. Rather, they are subject to examination and testing; each experiment should define its own rules and the results’ evaluation can judge which rules are better. Due to the novelty of our architecture, we restricted ourselves to the minimum *transitivity*; the following rules are considered:

1. The *synonyms* of a given *user tag* will be added as *system tags*.

2. The *similar terms* of a given *user tag* will be added as *system tags*.
3. The *related terms* of a given *user tag* will be added as *system tags*.
4. The *first level of hypernyms* of a given *user tag* will be added as *system tag*.
5. The translation, the translation synonyms, the translation related terms, and the translation similar terms of a given *user tag* will be added as *system tag*.

### 5.2.2 Social resources

As discussed in the previous chapter, the *clustering component* of the generic architecture for tagging systems addresses the challenge of *shorthand writing*. Two different scenarios for *clustering component* were suggested; either to perform the clustering process for real tag-based system(s), or to use the ready clusters offered by existing tag-based systems. In our work, the latter scenario is considered. In order to address the challenge of *shorthand writing*, the Flickr tagging system is considered to be the social resource from where the tag clusters will be obtained.

When the Flickr tagging system database is queried for the clusters of a given tag, a variable number of clusters will be retrieved with different number of tags in each cluster. According to our observation, most of the tags retrieve (have) only one cluster when submitted to Flickr database. Furthermore, the most related tags for a given tag in the retrieved clusters are the *Top-N* tags in the first cluster. Therefore, we decided to add the *Top-3* tags from the first cluster as *system tags*. Anyhow, this number might be changed in the *future work* based on the results that will be obtained and evaluated in the following chapters. Figure 5.3 shows the scope of our prototype.

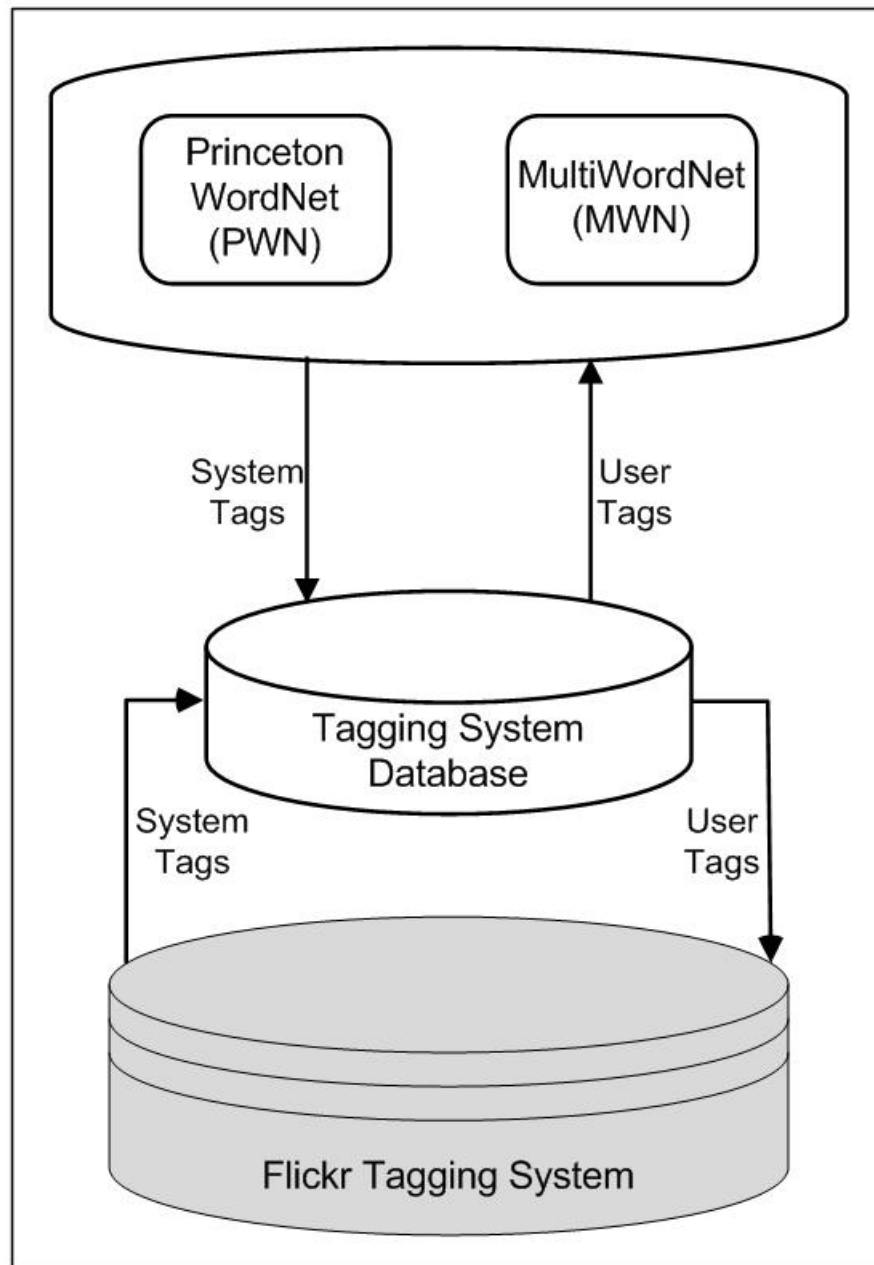


Figure 5.3: The scope of our prototype.

As seen in Figure 5.3, the prototype contains three components of the generic architecture. It contains the *semantic component* (in the top of the figure), the *clustering component* (in the bottom of the figure), and the *database component* (in the middle of the figure). The details of the *database component* in our prototype is discussed in Section 5.3.2.

## 5.3 Our algorithm for adding system tags

Two different scenarios for the *semantic component* were suggested in the previous chapter. These scenarios regard to the time of adding the *system tags*; either to add them at *tagging time* or at *searching time*. In our work, the former scenario is considered with some customisation since we are not operating a real tag-based system. Rather, we use sample data from a real tag-based system. The sample data is taken from YouTube, stored in a private database, and then our algorithm is applied on the private database.

### 5.3.1 YouTube

Since its debut in 2005, YouTube has become an extremely popular online video-sharing service. Registered users upload variety types of videos, with the exception of videos that are offensive or illegal, to the YouTube server for free [113]. Social communities create and annotate the content of YouTube by associating metadata that makes the videos searchable, and thus, accessible and survivable. Such metadata includes tags, category, brief description, thumbnails, and title.

What makes YouTube significant and popular for a wide range of people is the ease of watching and sharing videos in a conventional way for any Internet user, plus being totally free. The YouTube site offers an API (YouTube Data API for Java) that allows developers to build applications that can interact with the contents (e.g. upload video, annotate video, retrieve video information) [113].

The availability and ease of use of the YouTube Data API for Java was one of the reasons behind considering YouTube as the tag-based system from where we import the sample data. In addition, we consider the popularity of YouTube, and hence the acceptance and familiarity for users as our online experiment will be distributed to public users

through mailing lists (the online experiment is built on top of the sample data).

### 5.3.2 Database design

The design of our database is based on the one suggested in the *database component* of the generic architecture (see 4.3.5). YouTube database contains vast amount of data about videos and users. Our concern here is few data about the videos involving the user tags. Figure 5.4 shows a diagrammatic representation of the `Video` entity and its attributes that we need in our work.

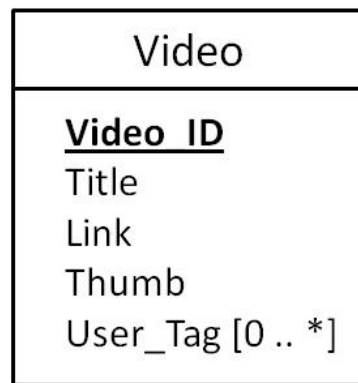


Figure 5.4: Diagrammatic representation of the *video* entity and its attributes.

The *primary key* for the entity is the `Video_ID`. The `Title`, the `Link`, and the `Thumb` attributes are imported to be used in the online experiment; the `Title` and the `Thumb` will be displayed to the end users whilst the `Link` will be used for hyper-linking the displayed `Title` and `Thumb` with the video URL on the YouTube website. These four attributes (`Video_ID`, `Title`, `Link`, and `Thumb`) are *single-valued* attributes while the `User_Tag` attribute is a *multi-valued* attribute.

*Single-valued* attribute is “the attribute that holds a single value for each occurrence of an entity type” [114]. In other words, each occurrence of the `Video` entity type has a maximum of one value, for example, the `Title` attribute (a video titled “*Sad Violin*”

cannot have another title).

*Multi-valued* attribute is “the attribute that holds multiple values for each occurrence of an entity type” [114]. For example, each occurrence of the entity type `Video` can have more than one value for the multi-valued attribute `User_Tag`. The notation “`User_Tag[0 .. *]`” implies that the `User_Tag` attribute can have no value (0) or more than one value (\*). In this case, a new relation should be created to represent the multi-valued attribute. The new relation should include the primary key of the original entity to act as a foreign key in the new relation [114]. Consequently, the `Video` relation will be two tables as shown in Figure 5.5.

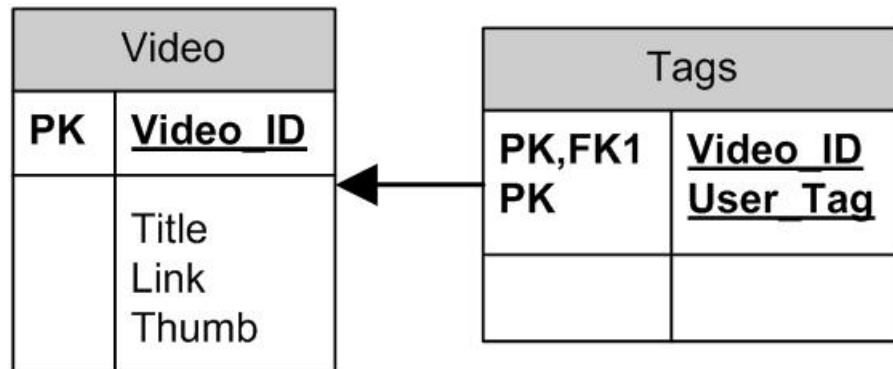


Figure 5.5: Logical diagram for the *video* entity tables.

In addition to these two tables, we created a third table to store the *system tags*. The attributes of the third table are: `User_Tag`, `System_Tag`, and `Tag_Type`. Each unique `User_Tag` in this table will have one or more `System_Tag` added from either semantic or social resources (the attribute `Tag_Type` specifies the resource of the `System_Tag`). The logical diagram for the three tables is shown in Figure 5.6.

The arrow between the `Videos` and `Tags_Master` tables refers to the referential integrity constraint (Foreign Key) between them, whilst it is not the case for the other arrow between the `Tags_Master` and `Tags_Detail` tables. In the latter case, the `User_`

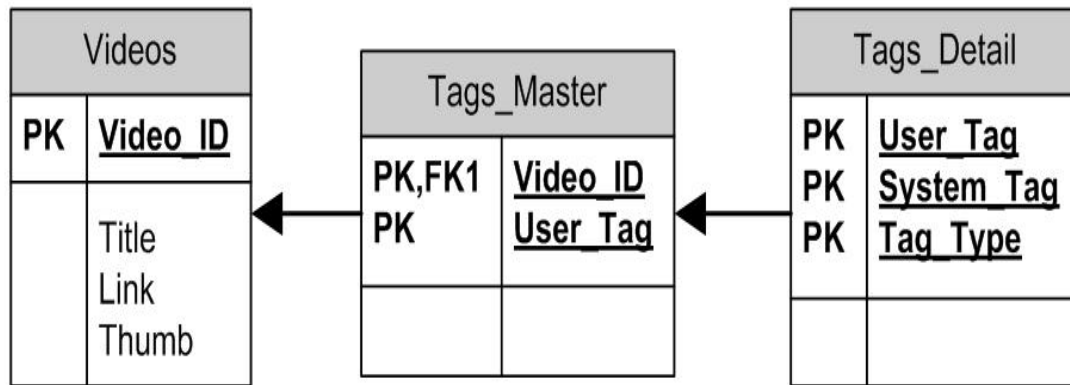


Figure 5.6: Logical diagram for our videos database.

Tag attribute in the Tags\_Detail table cannot reference the User\_Tag attribute in the Tags\_Master table as the latter is not a primary key. Therefore, a CHECK constraint is defined in the Tags\_Detail table to represent the relation between the Tags\_Detail and Tags\_Master tables. The constraint is defined as follows:

```
CHECK (User_Tag in (SELECT DISTINCT User_Tag
                    FROM Tags_Master))
```

### 5.3.3 Algorithm implementation

Our database is populated with data from three different resources; sample data and user tags from YouTube database, system tags from semantic ontologies (PWN and MWN), and system tags from Flickr clusters. To access these resources, import the required data, and store it in our database, we used the Java programming language.

Java is an Object-Oriented Programming (OOP) language that belongs to the third-generation (high-level) languages which makes it programmer-friendly with less low level facilities. The reason behind our choice to use Java is the availability of Java APIs to ac-



cess the abovementioned resources. The APIs are YouTube Data API<sup>4</sup>, Java API for WordNet Searching (JAWS)<sup>5</sup>, and Flickr Java API (flickrj)<sup>6</sup>. These APIs provide a set of callable methods and functions that enable developers to create client applications to upload, update, and retrieve data from YouTube, WordNet, and Flickr, respectively.

Our Java code consists of one class that contains several callable procedures to minimise the implementation dependency. We have two main types of procedures; **methods** that are called directly from the *main()* method, and **functions** that are called from the methods (not called by the *main()* method). The **methods do not return value** and are used to insert the system tags into our database, whereas the **functions return value** and do not insert system tags into our database.

Figure 5.7 shows a diagram for all the *methods* and *functions* used in our code. Each method is represented by one box with three subdivisions. The top subdivision shows the method/function *name*, the middle subdivision shows the *parameters*, and the bottom one is for the *returned value*. The arrows begin from the *calling* procedure and end in the *called* procedure.

As in Figure 5.7, the *main()* method has direct connection with seven methods (clock-wise in the figure: *it\_synonyms\_related\_similar()*, *en\_synonyms\_related\_similar()*, *en\_translation\_related\_similar()*, *it\_translation\_related\_similar()*, *tag\_clustering()*, *it\_hyponyms()*, and *en\_hyponyms()*). These methods are responsible for finding the relevant *system tags* and inserting them into our database. Depending on the resource of the *system tags*, finding the relevant *system tags* is done either inside these methods or by

---

<sup>4</sup>The documentation for YouTube Data API for Java is available online at <http://code.google.com/apis/gdata/javadoc/overview-summary.html>.

<sup>5</sup>The documentation for JAWS WordNet API for Java is available online at <http://lyle.smu.edu/~tspell/jaws/doc/overview-summary.html>.

<sup>6</sup>The documentation for flickrj (Flickr API for Java) is available online at <http://flickrj.sourceforge.net/api/overview-summary.html>.

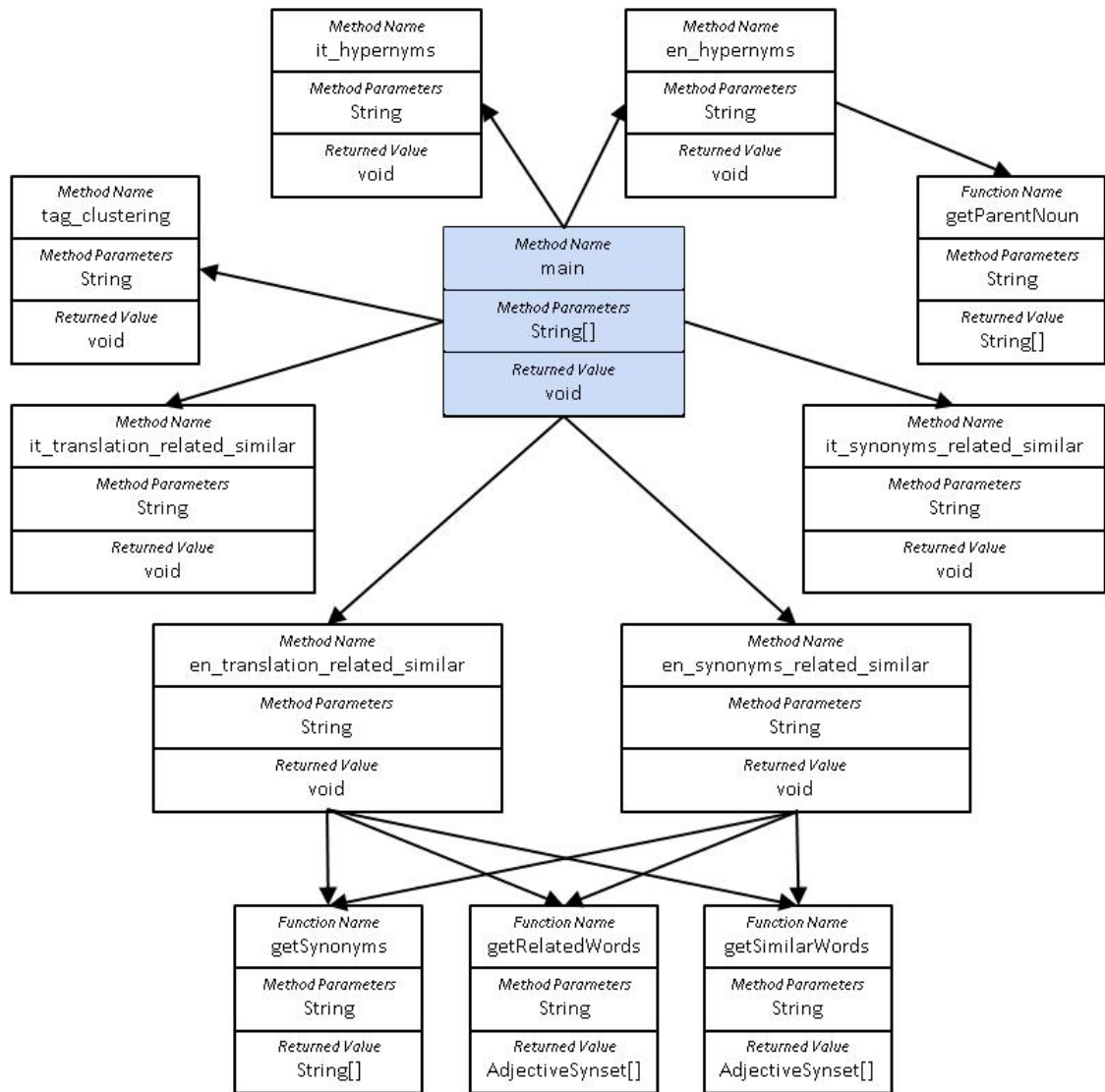


Figure 5.7: Methods diagram.

calling other functions (`getParentNoun()`, `getSimilarWords()`, `getRelatedWords()`, and `getSynonyms()`).

Before start calling any of these methods, the `main()` method queries the YouTube database using a set of English and Italian keywords via the YouTube Data API. This **initial**<sup>7</sup> set of keywords is stored in a `String` Array as follows:

<sup>7</sup>The database is initially populated using these keywords for piloting (testing) purposes. More elaboration about the “*pilot study*” is discussed in the next chapter.

```
// ----- //
// ** The English Keywords **
// ----- //
"education", "tutorial", "research", "student", "academy", "learning",
"technology", "system", "computer", "computing", "programming", "web",
"internet", "software", "engineering", "science", "media", "video", "tv",
"show", "music", "audio", "news", "cinema", "movie", "radio", "photo",
"ad", "advertisement", "entertainment", "comedy", "style", "model", "art",
"design", "beautiful", "paint", "beauty", "transportation", "car", "plane",
"train", "flight", "travel", "tourism", "holiday", "human", "people", "man",
"girl", "kid", "baby", "creature", "children", "arab", "social", "culture",
"religion", "history", "dancing", "sport", "football", "game", "business",
"product", "company", "money", "economy", "office", "mobile", "language",
"nature", "animal", "bird", "fish", "mammal", "jungle", "life", "world",
"health", "hospital", "military", "accommodation", "law", "utility", "event",
"funny", "sad", "communication", "food", "drink", "dish", "restaurant",
"beverage", "sex",
// ----- //
// ** The Italian Keywords **
// ----- //
"educazione", "tutorial", "ricerca", "studente", "accademia", "apprendimento",
"lezione", "Tecnologia", "sistema", "informatica", "programmazione", "scienza",
"musica", "pubblicita", "intrattenimento", "commedia", "foto", "stile",
"modello", "piano", "sociale", "storia", "mammifero", "legge", "beverage",
"bellezza", "arte", "disegno", "bello", "dipingere", "trasporto", "auto",
"treno", "volo", "viaggio", "turismo", "vacanza", "umano", "Persone",
"uomo", "ragazza", "creatura", "figli", "arabi", "cultura", "religione",
"ballare", "affari", "prodotto", "societa", "denaro", "economia", "ufficio",
"comunicazione", "linguaggio", "natura", "animale", "uccello", "pesce",
"giungla", "vita", "mondo", "salute", "ospedale", "militari", "alloggio",
"evento", "divertente", "triste", "cibo", "bere", "piatto", "ristorante",
"sessu"
```

By performing a loop throughout the `String Array`, the YouTube database is queried using one keyword in each iteration. We fixed the maximum number of videos to be retrieved for each keyword to “30”. After having iterated throughout all the keywords, all the videos retrieved from the YouTube are saved in one place (list) called “*video\_list*”.

For each video in the “*video\_list*”, we save the video information (the single-valued attributes such as `Video_ID`, `Title`, `Link`, and `Thumb`) in the `Videos` table in our database. Since each video has more than one tag (multi-valued attribute), another loop is performed over the video *user tags*. In each iteration, the `Video_ID` and the `User-Tag` are saved in the `Tags_Master` table, and the relevant *system tags* are added to the `Tags_Detail` table based on the `User-Tag` type; English and Italian word<sup>8</sup>, English word only, Italian word only, or shorthand writing word.

---

<sup>8</sup>Some words are English and Italian at the same time (e.g “*film*”).

The pseudo code of the *main()* method in Algorithm 5.3.1 illustrates how the videos' information retrieved from YouTube, and the *system tags* obtained from PWN, MWN, and Flickr are processed and saved in our database.

**Algorithm 5.3.1:** MAIN\_METHOD()

```

video_list ← getVideos()
for each video ∈ video_list
    {
        video_id ← getVideoId()
        video_title ← getVideoTitle()
        video_link ← getVideoURL()
        video_thumb ← getVideoThumb()
        Insert (video_id, video_title, video_link, video_thumb)
        Into Videos table
        video_tags_list ← getVideoTags()
        for each tag ∈ video_tags_list
            {
                Insert (Video_ID, User_Tag)
                Into Tags_Master table
                if current tag is English word and Italian word at the same time
                    {
                        EN_SYNONYMS_RELATED_SIMILAR(tag)
                        IT_SYNONYMS_RELATED_SIMILAR(tag)
                        EN_HYPERNYMS(tag)
                        IT_HYPERNYMS(tag)
                        IT_TRANSLATION_RELATED_SIMILAR(tag)
                        EN_TRANSLATION_RELATED_SIMILAR(tag)
                    }
                do {
                    if current tag is English word only
                        {
                            EN_SYNONYMS_RELATED_SIMILAR(tag)
                            EN_HYPERNYMS(tag)
                            IT_TRANSLATION_RELATED_SIMILAR(tag)
                        }
                    then {
                        if current tag is Italian word only
                            {
                                IT_SYNONYMS_RELATED_SIMILAR(tag)
                                IT_HYPERNYMS(tag)
                                EN_TRANSLATION_RELATED_SIMILAR(tag)
                            }
                        then {
                            EN_SYNONYMS_RELATED_SIMILAR(tag)
                            IT_TRANSLATION_RELATED_SIMILAR(tag)
                        }
                    }
                    else TAG_CLUSTERING(tag)
                }
            }
        }
    }

```

As abovementioned, the *main()* method has direct interaction with the **methods** rather than the **functions**. The methods<sup>9</sup> obtain the system tags from either the MWN or the

---

<sup>9</sup>Methods' names are uppercase-letter words separated by underscores.

Flickr clusters, using SQL statement or flickrj API respectively, and save the system tags in `Tags_Detail` table. The functions<sup>10</sup> obtain the system tags from PWN using the JAWS API and return them to the methods rather than saving them in the database.

The `EN_SYNONYMS_RELATED_SIMILAR()` method receives the *tag* as a `String` parameter, it finds the relevant English synonyms, related terms, and similar terms from the PWN via calling other functions (*getSynonyms()*, *getRelatedWords()*, and *getSimilarWords()* respectively), and it saves them in the `Tags_Detail` table as shown using the pseudo code notation in Algorithm 5.3.2.

**Algorithm 5.3.2:** `EN_SYNONYMS_RELATED_SIMILAR(tag)`

```

synonyms_list ← GETSYNONYMS(tag)
for each synonym ∈ synonyms_list
  do {
    Insert (tag, synonym, 'EN_SYNONYM')           (i)
    As (User_Tag, System_Tag, Tag_Type)
    Into Tags_Detail table;
  }
related_words_list ← GETRELATEDWORDS(tag)
for each related_word ∈ related_words_list
  do {
    Insert (tag, related_word, 'EN_RELATED')       (ii)
    As (User_Tag, System_Tag, Tag_Type)
    Into Tags_Detail table;
  }
similar_words_list ← GETSIMILARWORDS(tag)
for each similar_word ∈ similar_words_list
  do {
    Insert (tag, similar_word, 'EN_SIMILAR')       (iii)
    As (User_Tag, System_Tag, Tag_Type)
    Into Tags_Detail table;
  }

```

Each time we insert a new *system tag* into the `Tags_Detail` table, we insert three values into the table; the `User_Tag`, the `System_Tag`, and the `Tag_Type`. The `Tag_Type` is a description for the *system tag* (between two single quotations) as shown on lines (i), (ii), and (iii) in Algorithm 5.3.2.

---

<sup>10</sup>Functions' names are concatenated initially-capitalised words, except the first word "get".

The `EN_HYPERNYMS()` method receives the *tag* as a `String` parameter, it finds the relevant English hypernym from the PWN via calling the function `getParentNoun()`, and it saves them in the `Tags_Detail` table as shown in Algorithm 5.3.3.

**Algorithm 5.3.3:** `EN_HYPERNYMS(tag)`

```

hypernyms_list ← GETPARENTNOUN(tag)
for each hypernym ∈ hypernyms_list
  do { Insert (tag, hypernym, 'EN_HYPERNYM')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
  }
```

The *Italian relevant system tags* for the English *user tags* are inserted into the database using the method `IT_TRANSLATION_RELATED_SIMILAR()` (Algorithm 5.3.4).

**Algorithm 5.3.4:** `IT_TRANSLATION_RELATED_SIMILAR(tag)`

```

it_translation_&_synonyms_list ← (Select it_translation_&_synonyms
                                   From MWN)
for each it_translation_&_synonym ∈ it_translation_&_synonyms_list
  do { Insert (tag, it_translation_&_synonym, 'IT_TRANS_&_SYN')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
  }
it_related_words_list ← (Select it_related_words
                               From MWN)
for each it_related_word ∈ it_related_words_list
  do { Insert (tag, it_related_word, 'IT_RELATED')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
  }
it_similar_words_list ← (Select it_similar_words
                               From MWN)
for each it_similar_word ∈ it_similar_words_list
  do { Insert (tag, it_similar_word, 'IT_SIMILAR')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
  }
```

As shown in the previous algorithm, the *Italian relevant system tags* are: the Italian

translation, the Italian synonyms, the Italian related terms, and the Italian similar terms. They are obtained from the MWN using SQL statements. If the *user tag* is Italian word, then the *Italian relevant system tags* are also obtained from the MWN and inserted into our database using SQL statements as shown in Algorithm 5.3.5 and Algorithm 5.3.6.

**Algorithm 5.3.5:** IT\_SYNONYMS\_RELATED\_SIMILAR(*tag*)

```

synonyms_list ← (Select synonyms
                  From MWN)
for each synonym ∈ synonyms_list
  do { Insert (tag, synonym, 'IT_SYNONYM')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
      }
related_words_list ← (Select related_words
                       From MWN)
for each related_word ∈ related_words_list
  do { Insert (tag, related_word, 'IT_RELATED')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
      }
similar_words_list ← (Select similar_words
                        From MWN)
for each similar_word ∈ similar_words_list
  do { Insert (tag, similar_word, 'IT_SIMILAR')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
      }

```

**Algorithm 5.3.6:** IT\_HYPERNYMS(*tag*)

```

hypernyms_list ← (Select hypernyms
                    From MWN)
for each hypernym ∈ hypernyms_list
  do { Insert (tag, hypernym, 'IT_HYPERNYM')
        As (User_Tag, System_Tag, Tag_Type)
        Into Tags_Detail table;
      }

```

The *English relevant system tags* for the Italian *user tags* are inserted into the database using the method EN\_TRANSLATION\_RELATED\_SIMILAR() as shown in Algorithm

5.3.7.

**Algorithm 5.3.7:** EN\_TRANSLATION\_RELATED\_SIMILAR(*tag*)

```

en_translation ← (Select en_translation                                (i)
                    From MWN)
synonyms_list ← GETSYNONYMS(en_translation)                        (ii)
for each synonym ∈ synonyms_list
    do { Insert (tag, synonym, 'EN_TRANS-&_SYN')
          As (User_Tag, System_Tag, Tag_Type)
          Into Tags_Detail table;
    related_words_list ← GETRELATEDWORDS(en_translation)            (iii)
    for each related_word ∈ related_words_list
        do { Insert (tag, related_word, 'EN_RELATED')
              As (User_Tag, System_Tag, Tag_Type)
              Into Tags_Detail table;
    similar_words_list ← GETSIMILARWORDS(en_translation)            (iv)
    for each similar_word ∈ similar_words_list
        do { Insert (tag, similar_word, 'EN_SIMILAR')
              As (User_Tag, System_Tag, Tag_Type)
              Into Tags_Detail table;

```

The MWN contains two languages; English language and Italian language, whilst the PWN contains only the English language. In other words, we have two options from where to obtain the English *system tags* whereas we have only one resource from where to obtain the Italian *system tags*. For the English *system tags*, we tend to retrieve them from the PWN since it has a later version of the WordNet ontology. PWN contains the version 2.1 of the English WordNet ontology whilst MWN contains the version 1.6. The MWN is used for two purposes; retrieving the Italian *system tags*, and finding the corresponding translation for the *user tags*.

Back to Algorithm 5.3.7, we see that the corresponding English translation of the Italian tag is retrieved and saved in the variable *en\_translation* (on line (i)). Then the variable *en\_translation* is used to retrieve the English synonyms, related terms, and similar terms



from PWN via the functions *getSynonyms()*, *getRelatedWords()*, and *getSimilarWords()* respectively (on lines (ii), (iii), and (iv)).

The functions<sup>11</sup> that we used in our implementation employ the JAWS API's ready procedures to access the PWN and to retrieve the English relevant system tags. The pseudo code for these functions is shown below in Algorithms 5.3.8, 5.3.9, 5.3.10, and 5.3.11.

**Algorithm 5.3.8:** GETSYNONYMS(*tag*)

```
synonyms_list  $\leftarrow$  getSynset(tag)  
return (synonyms_list)
```

**Algorithm 5.3.9:** GETRELATEDWORDS(*tag*)

```
related_words_list  $\leftarrow$  getRelated(tag)  
return (related_words_list)
```

**Algorithm 5.3.10:** GETSIMILARWORDS(*tag*)

```
similar_words_list  $\leftarrow$  getSimilar(tag)  
return (similar_words_list)
```

**Algorithm 5.3.11:** GETPARENTNOUN(*tag*)

```
hypernyms_list  $\leftarrow$  getHypernyms(tag)  
return (hypernyms_list)
```

The TAG\_CLUSTERING() method shown in Algorithm 5.3.12 is used when the *user tag* is neither English nor Italian word; the *user tag* is *shorthand writing* tag. The method sends the *user tag* to the Flickr database using the flickrj API. In return, Flickr database

---

<sup>11</sup>The functions start with the word “*get*”, return values, do not insert data into our database, and called by the other methods rather than the *main()* method.

returns a set of clusters for the *user tag*. We pick the *top* – 3 tags from the first cluster and add them as *system tags* for the given *user tag*.

**Algorithm 5.3.12:** TAG\_CLUSTERING(*tag*)

*clusters\_list*  $\leftarrow$  getClusters(*tag*)

**comment:** The following loop has 1 iteration to get the first cluster only

**for** *i*  $\leftarrow$  0 to 1

**do** { **comment:** The following loop has 3 iterations to get the first 3 tags only

**for** *i*  $\leftarrow$  0 to 3

**do** { Insert (*tag*, *cluster\_tag*, 'CLUSTERING')

**do** { As (*User\_Tag*, *System\_Tag*, *Tag\_Type*)

Into *Tags\_Detail* table;

## 5.4 Summary

The prototype implemented the semantic component, the clustering component, and the database component of our generic architecture for tagging systems. These three components were selected since they comprise the main aspects of the generic architecture; the semantic aspect, the multilinguality aspect, and the clustering aspect.

The semantic aspect and the multilinguality aspects were covered by using the semantic ontologies PWN and MWN, respectively. System tags were extracted from these two ontologies based on semantic relations with the original user tags. The relations that link the system tags with their corresponding user tags are the *translation*, the *synonymy*, the *hypernymy*, the *similar terms*, and the *related terms* relations. The last aspect (clustering) was covered by using Flickr clusters as a source of system tags for the shorthand writing user tags.

The prototype implementation was accomplish by building a database (using MySQL

Database Management System (DBMS)), and applying our algorithm to fill the database with user tags and system tags. The YouTube website was used as a source for videos and their associated user tags. The system tags sources were semantic and social resources; PWN and MWN as semantic resources, and Flickr clusters as a social resource.

## Chapter 6

# Experiment: Rationale and Design

### *Objectives:*

---

- Introducing the experiment we conducted to test the prototype implementation presented in the previous chapter.
  - Presenting the graphical and functional design of the online environment we set as part of the experiment.
  - Presenting the database design for storing the collected data.
  - Discussing the sampling design and the pilot study.
-

## 6.1 Introduction

In the previous chapter, we introduced a prototype implementation to address some of the tagging systems challenges. Definitely, a measurement tool is needed in order to evaluate the efficiency of the proposed prototype. Therefore, we designed an experiment that collects the end-users' feedback regarding the enhancements that were carried out to improve the information retrieval process in the tag-based systems.

The experiment comprises an online environment to facilitate collecting the participants' opinions, and a database where the collected data is stored for later analysis. We provide a further discussion concerning the rationale behind the use of the experiment, and discuss how the experiment can support the testing of the defined hypotheses. Furthermore, the sampling design and the sample selection method are considered.

## 6.2 The experiment rationale

The nature of our research is social since we are investigating in the area of e-socialisation and online communities. All aspects of the Social Web are revolving around the user; the user is responsible for content generation, content classification, and metadata creation. Furthermore, the content is retrieved and consumed by users. Based on the users' feedback and satisfaction, some contents might be given a higher priority (e.g. viral videos on YouTube), and other contents might be marked as spam or even deleted (e.g. violent or repulsive content, hateful or abusive content, harmful dangerous act, etc).

In such a user-centric research area, resorting to *subjective approach* of research methods to evaluate our prototype is inevitable. We mean by the *subjective approach* that the data which will be analysed, for the prototype evaluation purpose, is constituted by ask-

ing people questions. The people from whom the data can be collected are usually called *subjects* or *population*. Rather than asking every member of the population, the data is collected from only a fraction of the population; the so-called *sample* [115]. This can be achieved by reviewing exact subjects' opinions.

Investigating the opinions of a sample of the population is a technique in which the needed information is systematically collected in an easier, quicker, less expensive, and more accurate way [116, 117].

According to [117], one of the basic reasons for organisations and individuals to make use of such investigation is the creation, or the modification, of a product or service they provide. In our case, we are modifying the *searching service* in tag-based systems. Therefore, conducting such investigation using an online environment is beneficial and can support our research aim.

The main purpose of the online environment, in our research, is to enable a convenient sample of end-users to search in our database, and then to collect their opinions about the relatedness between the retrieved results and the submitted search terms (keywords). Our database, as shown in the previous chapter, is populated with real data obtained from YouTube (including user tags). Furthermore, the database contains system tags, that are relevant to the real user tags, obtained from different semantic and social resources.

### **6.3 The experiment design and interface**

In our experiment, we used three design and programming technologies; HyperText Markup Language (HTML), JavaScript, and JavaServer Pages (JSP). JSP is used to handle the server-side connection with MySQL database (where our sample data is stored). JSP is

responsible for generating dynamic content from the database and HTML is responsible for the Web pages design in which the content is displayed. JavaScript is used for data validation and events reactions (e.g. a button is activated if a checkbox is ticked).

### 6.3.1 Introductory page

In the first page of our online environment, the user sees an introductory page that casts a glance at a general idea about our research, a general idea about the experiment, and a consent form. So that the participant can anticipate what (s)he will experience throughout the online environment.

To carry out the experiment, the user must consent to participate in the experiment. The button that enables the user to continue is *deactivated* unless the user *ticks* a checkbox as shown in Figure 6.1.

Once the participant pressed the button that indicates an agreement of the participation, (s)he will be taken into the search page.

### 6.3.2 Search page

The second page of the online environment is the page where users can type “*search keywords*” to be submitted to our database (MySQL database). The DBMS will search for matching records, and hence the related videos will be retrieved. The interface of this page is simple and designed in a *Google-like* fashion.

For each search trial, a maximum<sup>1</sup> of two different groups of videos will be retrieved (if there are related results); one group is for the videos that resulted from searching in

---

<sup>1</sup>The number of retrieved groups varies from 0 - 2 groups depending on the matching videos found in the database.

### About our research

This experiment is being conducted as part of a Ph.D research titled "Generic Architecture for Semantic Enhanced Tagging System". This research is being conducted by Murad Magableh (Ph.D Student), and supervised by Dr. Martin Ward and Dr. Antonio Cau, in the STRL Department at De Montfort University, Leicester, UK.

We investigate, whether or not, the semantic technology can improve the quality of information retrieval process in the social tag-based systems.

### About this survey

In this experiment, we used **YouTube** website for our experiment. All what you are kindly asked to do is:

1. Search using English or Italian languages only
2. The results will be one or two groups of **YouTube** videos
3. Evaluate whether these videos are related to your searching keywords or not
4. We use a scale from 1 - 6 to measure the relatedness
5. Submit your evaluation
6. You may repeat this process more than once.

### Consent form

Please note that:

- \* Your participation in this experiment is entirely voluntary.
- \* We do not collect any private information from participants; participants are anonymous.
- \* Only the "search keywords" and your rating will be collected. Yet, it is

☒ agree to participate in this survey

Continue




Figure 6.1: Our online environment - Introduction page.

the *user tags only*, and the other group is for the videos that resulted from searching in the *system tags only*. This difference between the two groups is not spelt out to the participants in order to avoid any bias in the their evaluation.

The participant is obliged to evaluate all the displayed results (videos). Therefore, the maximum number of results to be displayed for the submitted keywords can be predetermined by the participant. As seen in Figure 6.2, the user can determine how many videos will be displayed for each group in the results page. The default value for the videos for each group is *four* which means that a maximum total of *eight* videos will be displayed, and hence should be evaluated.

The search keywords submitted in the search page will search in the video tags. By definition, tags should be distinguishing keywords that are best describing objects. There-



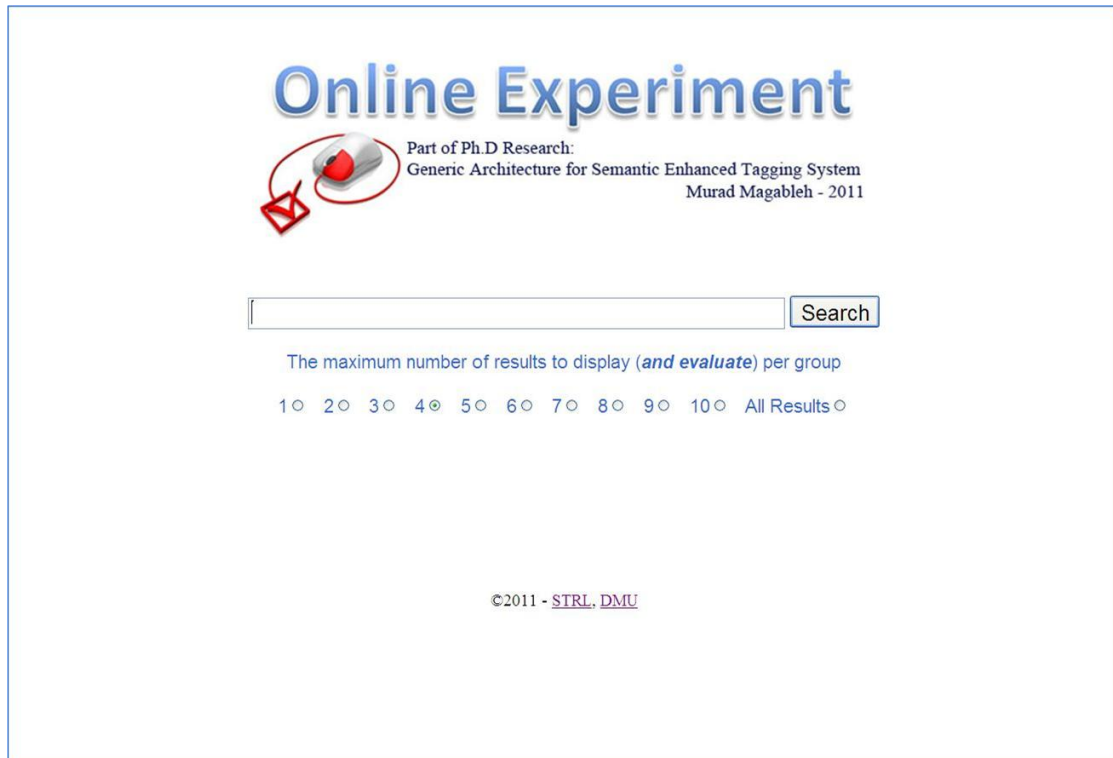


Figure 6.2: Our online environment - Search page.

fore, when users are tagging, they do not tend to use “*general words*” that cannot distinguish, or even describe, their content (e.g. “*the*”, “*a*”, “*of*”, etc). In YouTube, the user is not obliged to fill the tags field; it is optional to add tags. In case that the user does not insert any tag, the tags field will be filled by the same words used in the video title. For example, if the uploaded video title is “*The Lord of The Rings - The Return of The King*”, then the tags for this video will be “*The*”, “*Lord*”, “*of*”, “*Rings*”, “*Return*”, “*King*”<sup>2</sup>, unless the user specifies other tags. Therefore, many videos have such general words as tags.

If these words are used as tags in the search process, they will produce many videos that are not related. For example, if the searching keywords are “*The Lord of The Rings*”, the DBMS will be looking for any video that has any of these words as tags. So that any video that has the general words “*the*” or “*of*” will be retrieved which does not make

<sup>2</sup>No repetition in the tags.

sense. The words that we considered as general words belong to particular part of speech; *propositions* and *articles*. Indeed, not all the propositions and articles are very general; “*of*”, “*to*”, and “*in*” are general, whereas “*against*”, “*between*”, and “*outside*” are not. Therefore, we excluded some English propositions, English articles, Italian proposition, and Italian articles from the searching keywords, that what we so called “*non-searching keywords*”. The non-searching keywords are shown below.

```
// ----- //
                ** Non-Searching Keywords **
// ----- //
"of", "at", "in", "on", "for", "to", "with", "till", "by",
    "a", "an", "the", "as", "and", "or",
    "de", "del", "dello", "della", "dei", "degli", "delle",
    "a", "al", "allo", "alla", "ai", "agli", "alle",
    "da", "dal", "dallo", "dalla", "dai", "dagli", "dalle",
    "in", "nel", "nello", "nella", "nei", "negli", "nelle",
    "su", "sul", "sullo", "sulla", "sui", "sugli", "sulle",
    "con", "il", "lo", "la", "i", "gli", "le", "di"
```

In our online environment, any non-searching keyword will be excluded from the searching keywords, and hence, not submitted to the search engine.

Once the user types the searching keyword(s) and presses the “*search*” button shown in Figure 6.2, the results will be displayed in the results page.

### 6.3.3 Results page

The results page is divided horizontally into two columns, each column represents one group of results. The results retrieved by searching in the *user tags only* are displayed on the left hand-side column whereas the results retrieved by searching in the *system tags only* are displayed on the right hand-side column. The results page in Figure 6.3 shows the first four videos (in each group of results) retrieved using the keywords “*pretty girl*”.

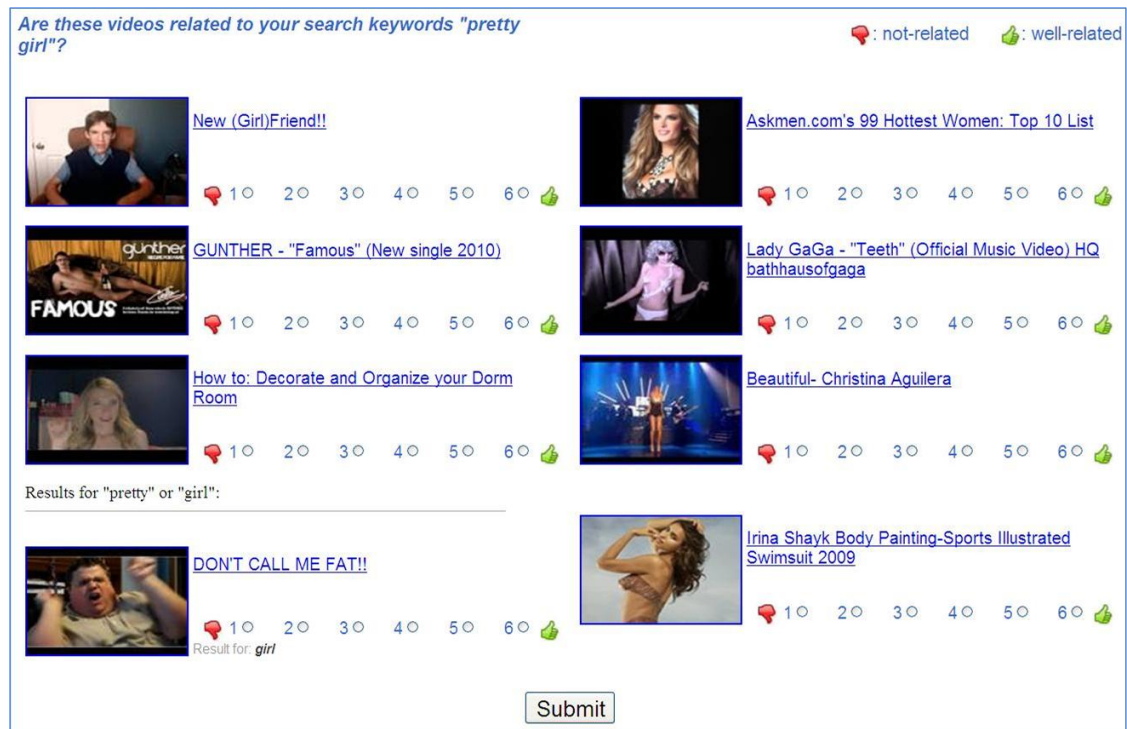


Figure 6.3: Our online environment - Results page.

Although the videos shown in the left hand-side column resulted from searching the original YouTube tags only, it is important to notice that submitting the same keywords in the example above to the YouTube website might retrieve different results. It is for the reason that YouTube considers other kinds of metadata, rather than the tags, that we do not consider in the search process.

As aforesaid, the second hypothesis H2 entails a comparison between the two groups of videos. For this comparison to be unbiased and fair-minded, we considered two issues in the videos evaluation; *blindness* and *fixed-conditions*.

By blindness we mean that the participant has no idea about the difference between the two groups of results. Therefore, (s)he will not try to give any socially accepted answers<sup>3</sup>.

<sup>3</sup>In the social research, participants might give some unreal answer (socially accepted answers) to please some parties.

By fixed-conditions we mean that all the factors that affect the users' evaluation of the two groups of results are identical. In other words, in each searching and evaluation trial, the following factors are considered:

- The two groups of results are produced from the *same* database; participants are not searching in two different samples of videos.
- The *same* participants are evaluating the two groups of results at the *same* time. This is important to avoid any variation that might happen due to the fact that humans' perceptions and attitudes might differ from time to time based on their moods, surrounding environment, and the time of the day, or of the week, in which they are evaluating the videos.
- The two groups of results are presented to the participants in *identical* modes and under the *same* conditions; both are given in a Web environment, both are presented using the *same* interface, and both are evaluated using the *same* scale.
- The two groups of results are produced for the *same* submitted keywords.
- The only *different* thing is that the kind of metadata, specifically tags, that is used to retrieve the results of each group is different.

As aforementioned, the layout of the results page shown in Figure 6.3 is divided horizontally into two columns. Each column is split into two sets of rows. Figure 6.4 illustrates the layout of the results page.

The left hand-side column displays the videos that have matching tags in the *user tags* to the submitted keywords in the `User_Tag` field in the `Tags_Master` table. Whereas the right hand-side column displays the videos that have matching tags to the submitted keywords in the `System_Tag` field in the `Tags_Detail` table<sup>4</sup>.

---

<sup>4</sup>See Figure 5.6 in Section 5.3.2.

Results retrieved by searching in the “ <b>User Tags</b> ” only using the “ <b>AND</b> ” logical operator between the searching keywords	Results retrieved by searching in the “ <b>System Tags</b> ” only using the “ <b>AND</b> ” logical operator between the searching keywords
Results retrieved by searching in the “ <b>User Tags</b> ” only using the “ <b>OR</b> ” logical operator between the searching keywords	Results retrieved by searching in the “ <b>System Tags</b> ” only using the “ <b>OR</b> ” logical operator between the searching keywords

Figure 6.4: The layout of the results page.

In the left hand-side column, the videos that have matching tags for *all* the submitted keywords are displayed first. Namely, an “**AND**” logical operator is used between the submitted keywords. For example, if the submitted keywords are “*pretty girl*”, then the videos that have the “*pretty*” keyword and the “*girl*” keyword as tags in the `User_Tag` field in the `Tags_Master` table will be shown in the top set of rows of this column. The select statement used to fill this partition is shown below.

```
select distinct v.video_id video_id_number, v.title video_title,  
               v.link video_link, v.thumb video_thumb  
from videos v, tags_master tm  
where  
  v.video_id = tm.video_id and  
  tm.user_tag = 'pretty girl'  
UNION
```

```
select distinct v.video_id video_id_number, v.title video_title,
               v.link video_link, v.thumb video_thumb
from videos v, tags_master tm0, tags_master tm1
where
  v.video_id = tm0.video_id and
  (
    tm0.user_tag = 'pretty' or
    tm0.user_tag like '%-pretty' or
    tm0.user_tag like 'pretty-%' or
    tm0.user_tag like '%-pretty-%' or
    tm0.user_tag like '% pretty' or
    tm0.user_tag like 'pretty %' or
    tm0.user_tag like '% pretty %' or
    tm0.user_tag like '%/_pretty' escape '/' or
    tm0.user_tag like 'pretty/_%' escape '/' or
    tm0.user_tag like '%/_pretty/_%' escape '/'
  )
and
  v.video_id = tm1.video_id and
  (
    tm1.user_tag = 'girl' or
    tm1.user_tag like '%-girl' or
    tm1.user_tag like 'girl-%' or
    tm1.user_tag like '%-girl-%' or
    tm1.user_tag like '% girl' or
    tm1.user_tag like 'girl %' or
    tm1.user_tag like '% girl %' or
    tm1.user_tag like '%/_girl' escape '/' or
    tm1.user_tag like 'girl/_%' escape '/' or
    tm1.user_tag like '%/_girl/_%' escape '/'
  )
)
```

After filling the top rows of the left hand-side column by the results retrieved using the previous select statement (if any), the bottom rows will be filled by the videos that have matching tags for *any* of the submitted keywords. Namely, an “**OR**” logical operator is used between the submitted keywords. In the previous example, the videos that have the “*pretty*” or the “*girl*” keywords, but not both of them, as tags in the User\_Tag field in the Tags\_Master table will be shown in the bottom rows of the left hand-side column. The select statement used to fill these rows is shown below.

```
select v.video_id video_id_number, v.title video_title,
       v.link video_link, v.thumb video_thumb,
       count(distinct case ----- (i)
when tm.user_tag = 'pretty' or
tm.user_tag like '%-pretty' or
```

```
tm.user_tag like 'pretty-%' or
tm.user_tag like '%-pretty-%' or
tm.user_tag like '% pretty' or
tm.user_tag like 'pretty %' or
tm.user_tag like '% pretty %' or
tm.user_tag like '%/_pretty' escape '/' or
tm.user_tag like 'pretty/_%' escape '/' or
tm.user_tag like '%/_pretty/_%' escape '/' then 'pretty'
when tm.user_tag = 'girl' or
tm.user_tag like '%-girl' or
tm.user_tag like 'girl-%' or
tm.user_tag like '%-girl-%' or
tm.user_tag like '% girl' or
tm.user_tag like 'girl %' or
tm.user_tag like '% girl %' or
tm.user_tag like '%/_girl' escape '/' or
tm.user_tag like 'girl/_%' escape '/' or
tm.user_tag like '%/_girl/_%' escape '/' then 'girl' end ) video_rank
from videos v, tags_master tm
where
v.video_id = tm.video_id and
(
tm.user_tag = 'pretty' or
tm.user_tag like '%-pretty' or
tm.user_tag like 'pretty-%' or
tm.user_tag like '%-pretty-%' or
tm.user_tag like '% pretty' or
tm.user_tag like 'pretty %' or
tm.user_tag like '% pretty %' or
tm.user_tag like '%/_pretty' escape '/' or
tm.user_tag like 'pretty/_%' escape '/' or
tm.user_tag like '%/_pretty/_%' escape '/' or
tm.user_tag = 'girl' or
tm.user_tag like '%-girl' or
tm.user_tag like 'girl-%' or
tm.user_tag like '%-girl-%' or
tm.user_tag like '% girl' or
tm.user_tag like 'girl %' or
tm.user_tag like '% girl %' or
tm.user_tag like '%/_girl' escape '/' or
tm.user_tag like 'girl/_%' escape '/' or
tm.user_tag like '%/_girl/_%' escape '/'
)
and
v.video_id not in ----- (ii)
(
select distinct v.video_id video_id_number, v.title video_title,
v.link video_link, v.thumb video_thumb
from videos v, tags_master tm0, tags_master tm1
where
v.video_id = tm0.video_id and
(
tm0.user_tag = 'pretty' or
```

```
tm0.user_tag like '%-pretty' or
tm0.user_tag like 'pretty-%' or
tm0.user_tag like '%-pretty-%' or
tm0.user_tag like '% pretty' or
tm0.user_tag like 'pretty %' or
tm0.user_tag like '% pretty %' or
tm0.user_tag like '%/_pretty' escape '/' or
tm0.user_tag like 'pretty/_%' escape '/' or
tm0.user_tag like '%/_pretty/_%' escape '/'
)
and
v.video_id = tm1.video_id and
(
  tm1.user_tag = 'girl' or
  tm1.user_tag like '%-girl' or
  tm1.user_tag like 'girl-%' or
  tm1.user_tag like '%-girl-%' or
  tm1.user_tag like '% girl' or
  tm1.user_tag like 'girl %' or
  tm1.user_tag like '% girl %' or
  tm1.user_tag like '%/_girl' escape '/' or
  tm1.user_tag like 'girl/_%' escape '/' or
  tm1.user_tag like '%/_girl/_%' escape '/'
)
)
group by video_id_number, video_title, video_link, video_thumb
order by video_rank desc ----- (iii)
```

To avoid duplication of videos in the left hand-side column of results, the latter select statement excludes the videos that were retrieved in the former one. Specifically, the videos that were retrieved and displayed in the top rows of the left hand-side column will not be retrieved and displayed again in the bottom rows of the same column (see line **(ii)**).

The display order of the videos in the bottom rows of the left hand-side column is tackled by adding the `count` function and the case statement on line **(i)**. The `video_rank` field is used to give a rank for each retrieved video based on the number of distinct matching tags (on line **(iii)**). The use of the `distinct` keyword in this select statement is to handle the case where a particular video has multiple similar (not identical) tags.



To explain this case, let us take, for example, a case where there are two videos (“*Video 1*” and “*Video 2*”) where “*Video 1*” has the tags “*girl*”, “*fat girl*”, and “*funny girl*”<sup>5</sup>. “*Video 2*” has the tags “*girl*” and “*pretty*”. If the search keywords are “*pretty young girl*”, the videos that have any combination of these tags will be displayed in the bottom rows of the left hand-side column. Namely, the videos that have any two of these tags (e.g. (“*pretty*”, “*young*”), (“*pretty*”, “*girl*”), (“*young*”, “*girl*”)), and the videos that have any one of these tags (e.g. “*pretty*”, “*young*”, “*girl*”). Rationally, the videos that have two matching keywords are expected to appear before the videos that have only one matching keyword. In this example, if we count the number of matching tags in “*Video 1*” and “*Video 2*”, we find that “*Video 1*” has three matching tags (three tags has the keyword “*girl*”) whilst “*Video 2*” has only two. Consequently, “*Video 1*” will appear ahead of “*Video 2*”. Indeed, “*Video 1*” does not have more matching keywords, but it has three repetition of the same tag “*girl*”. Therefore, counting the distinct occurrences of the tags will give a better display order of the results, and thus “*Video 2*” will appear ahead of “*Video 1*”.

In the right hand-side column, the videos that have matching tags in the *system tags* for *all* the submitted keywords are displayed first. Namely, an “**AND**” logical operator is used between the submitted keywords. For example, if the submitted keywords are “*pretty girl*”, then the videos that have the “*pretty*” keyword and the “*girl*” keyword as tags in the `System.Tag` field in the `Tags_Detail` table will be shown in the top rows of this column. The select statement used to fill these rows is shown below.

```
select distinct v.video_id video_id_number, v.title video_title,
               v.link video_link, v.thumb video_thumb
from videos v, tags_master tm, tags_detail td
where
  v.video_id = tm.video_id and
  tm.user_tag = td.user_tag and
  td.user_tag <> td.system_tag and ----- (i)
```

---

<sup>5</sup>YouTube allows users to add multiple-words tags, whereas some other tagging system (e.g. Del.icio.us) consider each word as a single tag.

```
    td.system_tag = 'pretty girl'
UNION
select distinct v.video_id video_id_number, v.title video_title,
               v.link video_link, v.thumb video_thumb
from videos v, tags_master tm0, tags_detail td0,
     tags_master tml, tags_detail tdl
where
v.video_id = tm0.video_id and
tm0.user_tag = td0.user_tag and
td0.user_tag <> td0.system_tag and ----- (ii)
(
    td0.system_tag = 'pretty' or
    td0.system_tag like '%-pretty' or
    td0.system_tag like 'pretty-%' or
    td0.system_tag like '%-pretty-%' or
    td0.system_tag like '% pretty' or
    td0.system_tag like 'pretty %' or
    td0.system_tag like '% pretty %' or
    td0.system_tag like '%/_pretty' escape '/' or
    td0.system_tag like 'pretty/_%' escape '/' or
    td0.system_tag like '%/_pretty/_%' escape '/'
)
and
v.video_id = tml.video_id and
tml.user_tag = tdl.user_tag and
tdl.user_tag <> tdl.system_tag and
(
    tdl.system_tag = 'girl' or
    tdl.system_tag like '%-girl' or
    tdl.system_tag like 'girl-%' or
    tdl.system_tag like '%-girl-%' or
    tdl.system_tag like '% girl' or
    tdl.system_tag like 'girl %' or
    tdl.system_tag like '% girl %' or
    tdl.system_tag like '%/_girl' escape '/' or
    tdl.system_tag like 'girl/_%' escape '/' or
    tdl.system_tag like '%/_girl/_%' escape '/'
)
```

In some cases, the added system tag is the same as the original user tag. This occurs when the hypernym or the translation of a particular word is the same as that word. Table 6.1 shows some examples from both Italian and English languages. To avoid retrieving the same videos in both of the results columns, the previous select statement excludes producing videos from the system tags that are identical to the user tags (on lines **(i)** and **(ii)**). Therefore, all the results that will appear in the right hand-side column are produced from the exclusively new system tags.

User Tag	System Tag	The Relation
scenario	scenario	Italian hypernym
scenario	scenario	Italian translation
pane	pane	Italian hypernym
album	album	Italian translation
bar	bar	Italian translation
barbecue	barbecue	Italian translation

Table 6.1: Some examples of similar *user tags* and *system tags*.

After filling the top rows of the right hand-side column by the results retrieved using the previous select statement (if any), the bottom rows will be filled by the videos that have matching tags for *any* of the submitted keywords. Namely, an “**OR**” logical operator is used between the submitted keywords. In the previous example, the videos that have the “*pretty*” or the “*girl*” keywords, but not both of them, as tags in the `System.Tag` field in the `Tags_Detail` table will be shown in the bottom rows of the right hand-side column. The select statement used to fill this partition is shown below.

```
select v.video_id video_id_number, v.title video_title,
       v.link video_link, v.thumb video_thumb,
       count(distinct case
when td.system_tag = 'pretty' or
td.system_tag like '%-pretty' or
td.system_tag like 'pretty-%' or
td.system_tag like '%-pretty-%' or
td.system_tag like '% pretty' or
td.system_tag like 'pretty %' or
td.system_tag like '% pretty %' or
td.system_tag like '%/_pretty' escape '/' or
td.system_tag like 'pretty/_%' escape '/' or
td.system_tag like '%/_pretty/_%' escape '/' then 'pretty'
when td.system_tag = 'girl' or
td.system_tag like '%-girl' or
td.system_tag like 'girl-%' or
td.system_tag like '%-girl-%' or
td.system_tag like '% girl' or
td.system_tag like 'girl %' or
td.system_tag like '% girl %' or
td.system_tag like '%/_girl' escape '/' or
td.system_tag like 'girl/_%' escape '/' or
td.system_tag like '%/_girl/_%' escape '/' then 'girl' end ) video_rank
```

```
from videos v, tags_master tm, tags_detail td
where
v.video_id = tm.video_id and
tm.user_tag = td.user_tag and
td.user_tag <> td.system_tag and
(
  td.system_tag = 'pretty' or
  td.system_tag like '%-pretty' or
  td.system_tag like 'pretty-%' or
  td.system_tag like '%-pretty-%' or
  td.system_tag like '% pretty' or
  td.system_tag like 'pretty %' or
  td.system_tag like '% pretty %' or
  td.system_tag like '%/_pretty' escape '/' or
  td.system_tag like 'pretty/_%' escape '/' or
  td.system_tag like '%/_pretty/_%' escape '/' or
  td.system_tag = 'girl' or
  td.system_tag like '%-girl' or
  td.system_tag like 'girl-%' or
  td.system_tag like '%-girl-%' or
  td.system_tag like '% girl' or
  td.system_tag like 'girl %' or
  td.system_tag like '% girl %' or
  td.system_tag like '%/_girl' escape '/' or
  td.system_tag like 'girl/_%' escape '/' or
  td.system_tag like '%/_girl/_%' escape '/'
)
and
v.video_id not in
(
select distinct v.video_id video_id_number
from videos v, tags_master tm0, tags_detail td0,
      tags_master tm1, tags_detail td1
where
  v.video_id = tm0.video_id and
  tm0.user_tag = td0.user_tag and
  td0.user_tag <> td0.system_tag and
  (
    td0.system_tag = 'pretty' or
    td0.system_tag like '%-pretty' or
    td0.system_tag like 'pretty-%' or
    td0.system_tag like '%-pretty-%' or
    td0.system_tag like '% pretty' or
    td0.system_tag like 'pretty %' or
    td0.system_tag like '% pretty %' or
    td0.system_tag like '%/_pretty' escape '/' or
    td0.system_tag like 'pretty/_%' escape '/' or
    td0.system_tag like '%/_pretty/_%' escape '/'
  )
and
  v.video_id = tm1.video_id and
  tm1.user_tag = td1.user_tag and
  td1.user_tag <> td1.system_tag and
```

```
(
  tdl.system_tag = 'girl' or
  tdl.system_tag like '%-girl' or
  tdl.system_tag like 'girl-%' or
  tdl.system_tag like '%-girl-%' or
  tdl.system_tag like '% girl' or
  tdl.system_tag like 'girl %' or
  tdl.system_tag like '% girl %' or
  tdl.system_tag like '%/_girl' escape '/' or
  tdl.system_tag like 'girl/_%' escape '/' or
  tdl.system_tag like '%/_girl/_%' escape '/'
)
)
group by video_id_number, video_title, video_link, video_thumb
order by video_rank desc
```

Each retrieved video, if any, in the aforementioned four sets of rows must be evaluated by the participant. The participant's decision of evaluating the retrieved videos might be based on the video title, video thumb, watching the video, or any combination of these. Therefore, misleading titles or thumbs might produce incorrect evaluation. The retrieved videos are evaluated using a 1-to-6 *Likert scale*.

### **Likert scale**

Scales are used to collect participants' responses and to compare these responses to each other. For facilitating the manipulation of the collected data, scales are coded with numbers in a systematic fashion. Using the scales can capture the participants' responses in a quick and accurate way [117].

One of the most common scales is the *Likert scale*, named for its creator, that is used for obtaining people's opinions and positions on certain issues. It is very popular among researchers due to its simplicity and flexibility in obtaining the respondents' degree of agreement or disagreement [117].

As seen in Figure 6.3, we use the Likert scale to collect the participants' opinions about the relatedness of the resulted videos to the submitted keywords. The scale we

used is a 1-to-6 rating scale. 1 expresses the participant's disagreement on the relatedness between the evaluated video and the searching keywords, whereas 6 expresses the participant's agreement.

There is a number of possible levels in the Likert scale; usually it would be a 1-to-5 scale [118]. Scales with levels less than 5 perform poorly, while scales with more level, up to 7, perform significantly better [119]. Being an odd-numbered or even-numbered is a considerable issue for the Likert scale. Odd-numbered scales have a middle value that indicates the neutrality of the respondent's degree of agreement or disagreement. To avoid the neutral or undecided choice, a forced-choice scale can be used just by using an even-numbered scale [118]. Accordingly, in our experiment an even-numbered scale of 6 levels was used to force participants to decide whether the video is related or not related, and to avoid carelessly answers.

Once the participant evaluated all the displayed videos using the Likert scale, (s)he will be transferred to the saving page.

### **6.3.4 Saving page**

The saving page has nothing to display except a message to let the participant know that the evaluation is done successfully as seen in Figure 6.5.

This page is shown for one second then it is automatically redirected to the "*Search Page*" (Figure 6.2), so the participant can have more search trials. Indeed, the saving page has more functions to do rather than content to display. In this page, all the participants' responses are collected and saved in the database for future analysis.

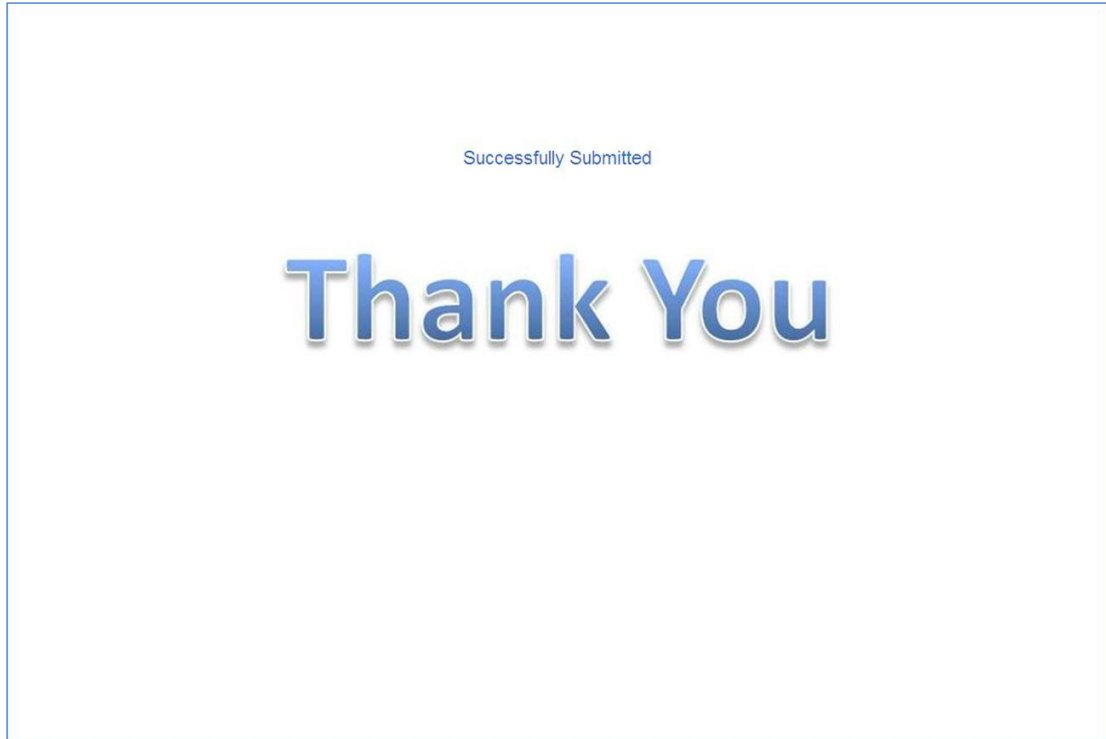


Figure 6.5: Our online environment - Saving page.

### 6.4 The database design

In the previous chapter<sup>6</sup>, we presented part of our database that is related to the sample data imported from the YouTube. Here we present the other part of our database which is concerned about the data collected using the online environment (e.g. the search terms used by the participants, the participants' evaluation, the number of videos retrieved in each search trial, etc).

There are some data about the participants' experience that should be captured while they are interacting with the online environment. If this data is not collected in a real-time basis, there will not be a chance to retrieve it. Therefore, we coded the online environment in such a manner that it collects any data which might be necessary in the analysis phase. The database design has to be consistent with the captured data.

---

<sup>6</sup>See 5.3.2.

Figure 6.6 shows the logical diagram for the database we used in our work; it includes the tables we used for the data collected in the online environment, in addition to the tables discussed in the previous chapter. Whenever a new participant makes a search trial, the system will generate a sequential number to be used as a trial identifier (`Trial_ID`). The `Trial_ID` and the keyword(s) used in that search trial (`Search_Keywords`) are stored in the `Search_Trials` table (shown on the top of Figure 6.6).

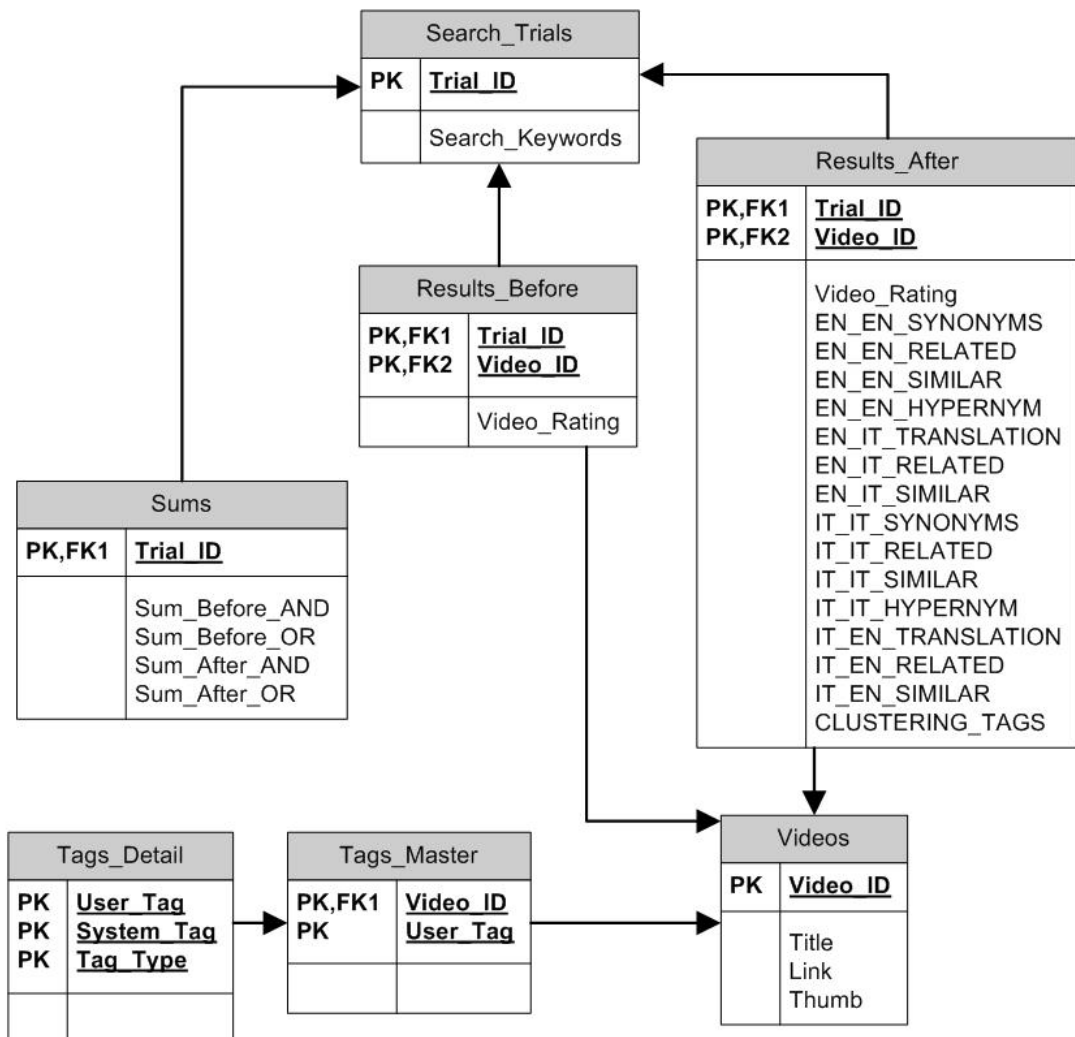


Figure 6.6: Logical diagram for our entire database (videos data + experiment data).

As aforesaid, the participant can select how many results to be displayed for each search trial; not all the retrieved videos are displayed. The real number of retrieved results in both groups are stored in the table `Sums`. The fields of the `Sums` table store the



summation of the retrieved videos (the displayed and the not displayed) in each part of the results page layout shown in Figure 6.4.

The respondents' rating of the displayed videos is stored in two tables based on the type of tags used to retrieve these videos. The rating of the videos that resulted from searching in the user tags only are saved in the table `Results_Before`, and the rating of the videos that resulted from searching in the system tags only are saved in the table `Results_After`. Both of these two tables have identical three fields; `Trial_ID`, `Video_ID`, and `Video_Rating`. The matching user tags that caused the evaluated video to be retrieved have one source only; they were written by the end user. This is not the case for the matching system tags that caused the evaluated video to be retrieved; these tags came from different sources (PWN, MWN, Flickr clusters) and different kinds of relations (synonymy, hypernymy, similar term, related terms, etc). For deeper and more accurate analysis of the efficiency of these system tags, we must record the source(s) of each system tag. Therefore, in the `Results_After` table, we added 15 fields that determine the sources of the system tags that caused the evaluated video to be retrieved.

The names of these fields have a regular notation<sup>7</sup>; the name consists of three syllables separated by underscores. The first syllable refers to the language of the corresponding user tag that caused the system tag to be added<sup>8</sup>. The second syllable refers to the language of the system tag itself, and the last syllable indicates the system tag's relation with the user tag (synonymy, hypernymy, translation, etc).

The values of these fields are either 0 or 1. For example, if the value of the field `EN_EN_SYNONYMS` is 1 and the values of the other 14 fields are 0s for a specific video, this means that the language of the system tag that caused the retrieval of that video is

---

<sup>7</sup>Except the last field `CLUSERING_TAGS`.

<sup>8</sup>"EN" refers to English and "IT" refers to Italian.

“*English*”, and that system tag was added to the database as a “*synonym*” of an original “*English*” user tag. For any retrieved video, at least one of these 15 columns will have a value of 1.

This specification of the system tag sources is significant in analysing the efficiency of each source independently. Therefore, we can compare between these sources to judge which one is more accurate and efficient in adding the system tags. For example, if the participants’ evaluation of the videos that have the value of 1 in the `EN_EN_SYNONYMS` field is higher than their evaluation of the videos that have the value of 1 in the `EN_EN_HYPERNYMY` field, this indicates that the former relation gives more accurate system tags than the latter relation. On the other hand, we can see which sources of the system tags caused the retrieval of the videos that had the best, as well as the worst, evaluation. For example, if the majority of videos that had 6 (well-related) in the users’ evaluation came from the source `EN_EN_SYNONYMS`, and the majority of videos that had 1 (not-related) in the users’ evaluation came from the source `EN_EN_HYPERNYM`, this means that the former relation gives more accurate system tags than the latter relation, and so on.

## 6.5 Sampling design

The principal idea in sampling is to extrapolate from the part to the whole; namely, from the “*sample*” to the “*population*” [120]. A *population* is an entire group of members that have at least one characteristic in common [12], while a *sample* is a subset of the population’s members [121]. Indeed, the size of the population from which the sample is drawn is seldom the determining factor and is largely irrelevant for the accuracy of the sample [120, 121]. Therefore, the population size is not our concern.

The methods for selecting the sample from the population is called the “*sampling de-*

*sign*” [120]. There are two main categories of sampling design techniques; *probability* and *non-probability*. The probability sampling is where each member in the population has an equal chance (probability) of being selected while in non-probability sampling some members have a greater chance of being selected than the others. Probability samples are preferable since they are more likely to produce representative samples. Yet, they are not appropriate in all cases. Different research aims and objectives require different methods of sampling [122]. The population in our experiment comprises any Internet user. In other words, the population in our case is widely dispersed. According to [121], non-probability sampling techniques are more appropriate as the population is so widely dispersed.

One of the most common sampling techniques is the *convenience sampling*. Convenience sampling is “a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher”. The subjects in convenience samples are the easiest to recruit for a study [123]. One of the main factors to consider when choosing the best sampling technique is the availability of time and resources [124]. For the researchers with restricted time and resources (e.g. students), convenience sampling is legitimately used and it is often the only feasible technique [125].

In addition to the limitations of time and resources, we used the convenience sampling technique in our study because of its ease, flexibility, less time-consuming, inexpensive-ness, and the availability of subjects [123, 124].

### **How big should a sample be?**

The number of subjects in a sample is a major concern in qualitative research. The sample size is one factor of representativeness of the population. Generally speaking, the larger the sample, the more representative of the population it is likely to be. One of the most

respected and preferable procedures to calculate the *minimum* required sample size for a study is known as the **power analysis** [126]. To estimate the sample size using the power analysis procedure, three components should be known or estimated by the researcher<sup>9</sup>.

The three factors are:

1. The significance criterion, *alpha* ( $\alpha$ ):  $\alpha$  is the risk of a Type I error, and it is standardised to be established to 0.05 [126, 127].
2. The effect size, *gamma* ( $\gamma$ ):  $\gamma$  is a measure of the extent to which the null hypothesis is false [122, 126]. Broadly, a small effect size is 0.20, a medium is 0.50, and a large is 0.80 [12, 122]. The effect size is usually considered from previous literature in the area of research. Revising the literature did not come across any studies that addressed effect size in similar studies. Consequently, in this study a conventional medium effect size was chosen.
3. The power ( $1 - \beta$ ):  $\beta$  is the risk of committing a Type II error. A conventional standard for  $\beta$  is established to 0.20. Therefore,  $1 - \beta$  is 0.80 [126].

There are different softwares available to calculate the sample size using the above-mentioned values. We performed the power analysis using the G-POWER [128]. Accordingly, the estimated sample size is 102 subjects. For more reliable findings and significant results, we included more subjects (102 is the minimum of subjects that can be included in the study not the exact number).

### 6.5.1 Piloting

Piloting refers to the practice of conducting a “*pilot study*” in research. A pilot study is a smaller-scale version, or a trial run, of a proposed study conducted in a preparation for

---

<sup>9</sup>Power analysis is a sophisticated method for sample size estimation, and it needs some statistical knowledge as a prerequisite. It requires familiarity with statistical concepts such as the **effect size**, **Type I error**, and **Type II error**. For more details see [126].

the major study. It is usually developed similarly to the proposed study to assess the feasibility of the full-scale study, and to test the adequacy of research instruments [122, 129].

As aforementioned in Section 5.3.3, our database was initially filled using a set of keywords. For each keyword, we set the maximum number of videos that can be retrieved from the YouTube to be 30. To make sure that the data imported from the YouTube is enough for conducting our experiment, we conducted a preliminary experiment using the same conditions and similar subjects. We received a general feedback from the subjects that there was a very limited number of videos, or no videos, retrieved for their searching keywords.

Analysing the approach in which we filled our database with YouTube videos, we realised the reason behind the feedback we received in the conducted pilot study. Although we imported 4,928 videos from the YouTube into our database, these videos belong to a restricted number of categories; the videos do not cover enough areas of interest for the potential subjects. The 169 English and Italian keywords used in Section 5.3.3 cover only 95 interests (the Italian keywords are the translation of the same English keywords). Because we imported 30 videos for each used keywords, the total number of retrieved videos is relatively high but the variety of categories is low; small number of categories with high number of videos in each category.

Therefore, we changed the approach in which we fill our database to cover more categories with small number of videos for each category. This required no changes to our algorithm; the change was only for the keywords and for the maximum number of videos that can be retrieved from the YouTube for each keyword. For each of the 1,163 keywords listed in Appendix A.1, a maximum of 7 videos were retrieved from the YouTube and stored in our database. Accordingly, our real experiment has been conducted on the latter

version of the data where the database contains 7,461 videos. More statistics about the data on which we conducted our experiment can be found in Appendix A.2.

### **6.6 Summary**

Since the user is the heart of the Social Web, a subjective approach was adopted in order to test the prototype implemented in the previous chapter. An online environment was set up to give the end-users the opportunity to search in our database and evaluate the retrieved results. Devoid of their awareness, users were evaluating results that were retrieved using two different types of tags; the old type (user tags), and our proposed system tags. The videos were evaluated using an even-numbered Likert scale of 6 levels ranging from not-related (1) to well-related (6).

The participants were selected using the convenience sampling technique which is one of the most commonly used sampling techniques. The sample size is, at least, 102 subjects, which was calculated using the power analysis procedure.

A pilot experiment was conducted before the main experiment in order to check, and improve if necessary, the design of our experiment. Consequently, the method used to import the YouTube was modified, and the database was refilled with new videos.

# Chapter 7

## Results and analysis

### *Objectives:*

---

- Discussing the techniques used for preparing the collected raw data for analysis.
  - Introducing the ideas of “*data collapsing*” and “*data dissection*”.
  - Describing the collected data using descriptive statistical measures.
  - Using inferential statistical procedures for drawing conclusions that can be generalised from our analysed sample to the population.
  - Discussing the results in the light of our research hypotheses.
-

## 7.1 Introduction

In the previous chapter, we discussed the experiment that was designed for testing our prototype. Namely, the experiment was for collecting the users' evaluation of the relatedness between the search keywords and the videos retrieved by using two groups of tags; user tags and system tags. After gathering the opinions of the selected sample, the raw data is stored in a MySQL database and is ready for manipulation and analysis.

This chapter will introduce the data preparation, manipulation, and analysis using Standard Query Language (SQL) statements, Microsoft Office Excel spreadsheet application, and Statistical Package for the Social Sciences (SPSS) application. Moreover, it will discuss the results in terms of two statistical techniques; *descriptive statistics* and *inferential statistics*.

For giving a better understanding of the efficiency of system tags, the results will be collapsed and dissected, then analysed again. The subjects' answers will be collapsed on a bipolar basis, and the collected data will be dissected based on the system tags sources using a variety of criteria.

## 7.2 Preparation of the data for analysis

Once the participants' answers have been collected (by any means), they are usually transformed into a data file that is appropriate for computer analysis [115]. This process is time consuming and tedious but essential and prerequisite for the data analysis [122].

In our case, the data was collected by computerised means and stored in a relational database tables (MySQL). Even though the data is in a digital format, it is not in ap-



propriate format for statistical analysis purposes. For analysing the collected data set, SPSS statistical package was used. In order to import the right data in the right format from the MySQL database into the SPSS application, two intermediary means were used *respectively*:

1. *HeidiSQL application*: HeidiSQL is a third-party application with graphical interface to facilitate MySQL database access and management. It was used to execute the SQL statements required to extract the right data from the tables and export the results into text files.
2. *Microsoft Office Excel spreadsheet application*: The Microsoft Office Excel was used to read the text files produced in the previous stage and organise the text files data in a grid of cells arranged in identified rows and columns. Organising the data this way makes it easy to read and manage by the SPSS.

As previously discussed in Section 6.4 and shown in Figure 6.6, the subject's evaluation is stored in two tables; `Results_Before` and `Results_After`. Each of the two tables has a field named `Video_Rating` that holds the value of the Likert scale item that was selected by the user. The `Results_Before` table stores the `Video_Rating` of the videos retrieved by searching in the user tags, whereas the `Results_After` table stores the `Video_Rating` of the videos retrieved by searching in the system tags.

Since we are comparing the rating of videos produced by searching in user tags with the rating of videos produced by searching in system tags, the following two select statements were used to retrieve these ratings:

- To retrieve the rating of the videos produced by searching in the user tags, we used this select statement:

```
SELECT Video_Rating
FROM Results_Before;
```

- To retrieve the rating of the videos produced by searching in the system tags, we used this select statement:

```
SELECT Video_Rating
FROM Results_After;
```

The previous two select statements produced two columns of data. The values of these two columns range from 1 to 6 (the Likert scale values). Afterwards, these two columns are transformed into a SPSS data file that is ready to be analysed.

### 7.2.1 Data collapsing

As previously discussed, the Likert scale used in this study is a 6-points scale. Collapsing Likert scales into fewer response categories is a commonly used technique in public opinion research [130]. Usually this is done by combining the *Positive* responses in one category and the *Negative* responses in another category to produce dichotomous categories. If the *Neutral* response is considered, trichotomous categories might be produced (e.g. see [131]).

Collapsing response categories produces fewer numbers which makes the data easier to comprehend [132]. Moreover, it helps capturing trends in data, and thus, facilitates inferences. This would improve the intelligibility of the analysis outcomes [133]. However, some information will be lost for the reader [132].

After analysing the original participants' responses using the 6 categories, the results were analysed from a different angle in order to explore another option of the results analysis. Therefore, the Likert scale results were collapsed into dichotomous categories; the Likert values range from 1 to 3 were collapsed to be 1 (*not-related*), and the Likert values range from 4 to 6 were collapsed to be 6 (*well-related*).

### 7.2.2 Data Dissection

Back to Figure 6.6, the `Results_After` table contains 15 fields. These fields refer to the source of the system tag that causes a specific video to be retrieved. The values of these fields are either 0 or 1. For example, if the value of the field `EN_EN_SYNONYMS` is 1 and the values of the other 14 field are 0s for a specific video, this means that the language of the system tag that caused the retrieval of that video is “*English*”, and that system tag was added to the database as a “*synonym*” of an original “*English*” user tag. For any retrieved video, at least one of these 15 columns will have a value of 1. Table 7.1 puts in plain words the differences among the 15 system tags sources:

Source Description	User Tag Language	System Tag Language	Relation
EN_EN_SYNONYMS	English	English	Synonymy
EN_EN_HYPERNYM	English	English	Hypernymy
EN_EN_SIMILAR	English	English	Similar terms
EN_EN_RELATED	English	English	Related terms
EN_IT_TRANSLATION	English	Italian	Translation
EN_IT_SIMILAR	English	Italian	Similar terms
EN_IT_RELATED	English	Italian	Related terms
IT_IT_SYNONYMS	Italian	Italian	Synonymy
IT_IT_HYPERNYM	Italian	Italian	Hypernymy
IT_IT_SIMILAR	Italian	Italian	Similar terms
IT_IT_RELATED	Italian	Italian	Related terms
IT_EN_TRANSLATION	Italian	English	Translation
IT_EN_SIMILAR	Italian	English	Similar terms
IT_EN_RELATED	Italian	English	Related terms
CLUSTERING_TAGS	Shorthand writing	N/A	The same cluster

Table 7.1: Explanation of the system tags sources and the language of their related user tags.

### Source-based dissection

In order to study the effect of each system tag source individually, we dissected the participants' evaluation of the videos retrieved by using system tags into 15 *subsets*; each subset contains the participants' evaluation of a group of videos. All videos in each group were retrieved by using system tags that came from one source of the 15 system tags sources.

Afterwards, each dissected *subset* was compared to the participants' evaluation of the videos retrieved by using user tags. For instance, the *subset* of participants' evaluation of the videos retrieved by using system tags that came from the EN\_EN\_SYNONYMS source was compared to the participants' evaluation of the videos retrieved by using user tags. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM    Results_After
WHERE   EN_EN_SYNONYMS = 1;
```

Similar 14 select statements were written for the rest of the 15 system tags sources, the only change in these select statements is the name of the field that appears in the WHERE clause (EN\_EN\_SYNONYMS); each time it is substituted with a different source name (EN\_EN\_HYPERNYM, EN\_EN\_SIMILAR, ..., etc). The produced 15 columns of data contain values that range from 1 to 6 (the Likert scale values). Afterwards, these columns are transformed into a SPSS data file that is ready to be analysed.

### Language-based dissection

System tags and user tags in this experiment belong to two languages; English and Italian. That is; a user tag and its relevant system tag might belong to the same language (e.g. both are English, or both are Italian), or they belong to different languages (e.g. one is

English and the other is Italian, and vice versa). In order to study the efficiency of system tags when their corresponding user tags belong to the same, or different, language, we dissected the system tags into 4 *subsets*, each *subset* was compared to the participants' evaluation of the videos retrieved by using user tags:

1. **The corresponding user tag and the system tag are both English (the source starts with “EN\_EN”)**: In this *subset*, we take the participants' evaluation of the videos retrieved by using system tags that came from the sources: **EN\_EN\_SYNONYMS**, **EN\_EN\_HYPERNYM**, **EN\_EN\_SIMILAR**, and **EN\_EN\_RELATED**. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM   Results_After
WHERE  EN_EN_SYNONYMS = 1 OR
       EN_EN_HYPERNYM = 1 OR
       EN_EN_SIMILAR  = 1 OR
       EN_EN_RELATED  = 1;
```

2. **The corresponding user tag and the system tag are both Italian (the source starts with “IT\_IT”)**: In this *subset*, we take the participants' evaluation of the videos retrieved by using system tags that came from the sources: **IT\_IT\_SYNONYMS**, **IT\_IT\_HYPERNYM**, **IT\_IT\_SIMILAR**, and **IT\_IT\_RELATED**. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM   Results_After
WHERE  IT_IT_SYNONYMS = 1 OR
       IT_IT_HYPERNYM = 1 OR
       IT_IT_SIMILAR  = 1 OR
       IT_IT_RELATED  = 1;
```

3. **The corresponding user tag is English and the system tag is Italian (the source starts with “EN\_IT”)**: In this *subset*, we take the participants’ evaluation of the videos retrieved by using system tags that came from the sources: **EN\_IT\_TRANSLATION**, **EN\_IT\_SIMILAR**, and **EN\_IT\_RELATED**. To retrieve this *subset* of participants’ evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM   Results_After
WHERE  EN_IT_TRANSLATION = 1 OR
       EN_IT_SIMILAR      = 1 OR
       EN_IT_RELATED      = 1;
```

4. **The corresponding user tag is Italian and the system tag is English (the source starts with “IT\_EN”)**: In this *subset*, we take the participants’ evaluation of the videos retrieved by using system tags that came from the sources: **IT\_EN\_TRANSLATION**, **IT\_EN\_SIMILAR**, and **IT\_EN\_RELATED**. To retrieve this *subset* of participants’ evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM   Results_After
WHERE  IT_EN_TRANSLATION = 1 OR
       IT_EN_SIMILAR      = 1 OR
       IT_EN_RELATED      = 1;
```

The produced 4 columns of the language-based dissection contain values that range from 1 to 6 (the Likert scale values). Afterwards, these columns are transformed into a SPSS data file that is ready to be analysed.

### **Relation-based dissection**

A system tag is added based on a semantic relation that relates it to an original user tag. The relation might be synonymy, hypernymy, similar terms, related terms, or translation.

In order to study the efficiency of system tags based on their relationship with user tags (regardless the tags language), we dissected the system tags into 5 *subsets*, each *subset* was compared to the participants' evaluation of the videos retrieved by using user tags:

1. **The relation between the user tag and the system tag is synonymy (the source ends with “SYNONYMS”)**: In this *subset*, we take the participants' evaluation of the videos retrieved by using system tags that came from the sources: EN\_EN\_**SYNONYMS** and IT\_IT\_**SYNONYMS**. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM Results_After
WHERE EN_EN_SYNONYMS = 1 OR
      IT_IT_SYNONYMS = 1;
```

2. **The relation between the user tag and the system tag is hypernymy (the source ends with “HYPERNYM”)**: In this *subset*, we take the participants' evaluation of the videos retrieved by using system tags that came from the sources: EN\_EN\_**HYPERNYM** and IT\_IT\_**HYPERNYM**. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM Results_After
WHERE EN_EN_HYPERNYM = 1 OR
      IT_IT_HYPERNYM = 1;
```

3. **The relation between the user tag and the system tag is similar (the source ends with “SIMILAR”)**: In this *subset*, we take the participants' evaluation of the videos retrieved by using system tags that came from the sources: EN\_EN\_**SIMILAR**, IT\_IT\_**SIMILAR**, EN\_IT\_**SIMILAR**, and IT\_EN\_**SIMILAR**. To retrieve this *subset* of participants' evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM Results_After
WHERE EN_EN_SIMILAR = 1 OR
      IT_IT_SIMILAR = 1 OR
      EN_IT_SIMILAR = 1 OR
      IT_EN_SIMILAR = 1;
```

4. **The relation between the user tag and the system tag is related (the source ends with “RELATED”)**: In this *subset*, we take the participants’ evaluation of the videos retrieved by using system tags that came from the sources: EN\_EN\_**RELATED**, IT\_–IT\_**RELATED**, EN\_IT\_**RELATED**, and IT\_EN\_**RELATED**. To retrieve this *subset* of participants’ evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM Results_After
WHERE EN_EN_RELATED = 1 OR
      IT_IT_RELATED = 1 OR
      EN_IT_RELATED = 1 OR
      IT_EN_RELATED = 1;
```

5. **The relation between the user tag and the system tag is translation (the source ends with “TRANSLATION”)**: In this *subset*, we take the participants’ evaluation of the videos retrieved by using system tags that came from the sources: EN\_–IT\_**TRANSLATION**, and IT\_EN\_**TRANSLATION**. To retrieve this *subset* of participants’ evaluation from the database, the following SQL statement was used:

```
SELECT Video_Rating
FROM Results_After
WHERE EN_IT_TRANSLATION = 1 OR
      IT_EN_TRANSLATION = 1;
```

The produced 5 columns of the relation-based dissection contain values that range from 1 to 6 (the Likert scale values). Afterwards, these columns are transformed into a SPSS data file that is ready to be analysed.



### 7.3 Descriptive statistics

Statistics are usually classified as either *descriptive* or *inferential* [126]. Descriptive statistics are used to summarise the collected data in a meaningful way. They provide simple summaries of large quantities of data using measures (e.g. mean, median, mode, frequencies) that are easily understood by observers. Descriptive statistics might include graphical and/or numerical techniques for showing concise summaries of data [12].

Indeed, descriptive statistics can describe only the sample under investigation, but they cannot draw conclusions that can be generalised to the population. Therefore, descriptive statistics are not used to make inferences regarding research hypotheses [134, 135].

In 204 search trials, the number of retrieved videos was 1,391 videos using either user tags or system tags; 704 videos were retrieved using user tags, whereas 687 videos were retrieved using system tags. Table 7.2 compares the *mean*<sup>1</sup>, the *median*<sup>2</sup>, and the *mode*<sup>3</sup> for the participants' evaluation of both groups of videos. The values of the three measures are almost the same; identical scores for the *median* and the *mode* with slightly different scores for the *mean*.

We can notice that the most frequently selected value (mode) in evaluating both of the groups was 1 (not-related). The main reasons behind this are:

1. The only searchable metadata in this experiment was restricted to be the tags (as mentioned earlier). In other words, not all the available metadata are used during the search. Conducting the same experiment under different conditions (considering more metadata during the search) will retrieve different videos, and thus, will give

---

<sup>1</sup>The mean (average) is the sum of all values divided by the number of values [12].

<sup>2</sup>The median (middle item) is the central value of an ordered list of values [12].

<sup>3</sup>The mode (modal) is the most frequently occurring value in a set of values [12].

	Participants' evaluation for videos retrieved using <b>User Tags</b>	Participants' evaluation for videos retrieved using <b>System Tags</b>
Number of videos	704	687
Mean	3.02	3.01
Median	2	2
Mode	1	1

Table 7.2: The descriptive measures for the whole data set.

different evaluation.

2. As aforesaid, the experiment was conducted on a subset of the videos published on YouTube website. Therefore, it was expected that several search terms will not find matching videos.

Nevertheless, the *second* most frequently selected value was 6 (well-related). Table 7.3 shows the frequencies of each Likert scale value selected in both groups of videos.

	Frequency of evaluations for videos retrieved using <b>User Tags</b>	Frequency of evaluations for videos retrieved using <b>System Tags</b>
1	308	318
2	67	57
3	50	31
4	46	46
5	48	49
6	185	186
Sum	704	687

Table 7.3: The frequencies of participants' evaluation for both groups.

It is notable from Table 7.3 that most of the videos' evaluation fall under the two extreme categories; either 1 (not-related) or 6 (well-related). The graphical representation of the numbers presented in Table 7.3 is shown in Figure B.2 and Figure B.3 in Appendix

B.1. Appendix B.1 provides further descriptive statistics about the whole data set.

The following two subsections show some descriptive statistics for the collapsed and dissected versions of data.

### 7.3.1 Collapsed data

After collapsing the data, the participants' evaluation values became either 1 or 6. Table 7.4 compares the *mean*, the *mediam*, and the *mode* of the *collapsed* participants' evaluation of both groups of videos. The values of the three measures are almost the same; identical scores for the median and the mode with slightly different scores for the mean. It is noteworthy that both of Table 7.2 (before collapsing) and table 7.4 (after collapsing) show the same similarity of the mean, median, and mode between the two groups of videos. More statistics about the collapsed data set can be found in Appendix B.2.

	Participants' evaluation for videos retrieved using <b>User Tags</b>	Participants' evaluation for videos retrieved using <b>System Tags</b>
Number of videos	704	687
Mean	2.98	3.05
Median	1	1
Mode	1	1

Table 7.4: The descriptive measures for the collapsed data set.

### 7.3.2 Dissected data

The three groups of dissected participants' evaluation for the videos retrieved using system tags, presented in Section 7.2.2, were compared to the participants' evaluation for the videos retrieved using user tags.

**Source-based dissection**

Table 7.5 compares the participants' evaluation for videos retrieved using user tags with the *source-based dissected* evaluation for videos retrieved using system tags in terms of the *number of videos*, the *mean*, the *mediam*, and the *mode*.

	Number of videos	Mean	Median	Mode
Participants' evaluation for videos retrieved using <b>User Tags</b>	704	3.02	2	1
EN_EN_SYNONYMS	251	2.88	1	1
EN_EN_HYPERNYM	153	2.75	1	1
EN_EN_SIMILAR	29	2.83	2	1
EN_EN_RELATED	7	3.43	3	1
EN_IT_TRANSLATION	52	3.33	3.5	1
EN_IT_SIMILAR	0	N/A	N/A	N/A
EN_IT_RELATED	0	N/A	N/A	N/A
IT_IT_SYNONYMS	48	3.38	3	1
IT_IT_HYPERNYM	5	4.40	4	3
IT_IT_SIMILAR	0	N/A	N/A	N/A
IT_IT_RELATED	0	N/A	N/A	N/A
IT_EN_TRANSLATION	16	2.69	1	1
IT_EN_SIMILAR	0	N/A	N/A	N/A
IT_EN_RELATED	0	N/A	N/A	N/A
CLUSTERING_TAGS	238	2.85	2	1

Table 7.5: The descriptive measures for the *source-based dissected* data set.

Table 7.5 shows irregularity in terms of the *number of videos* retrieved using the 15 system tags sources; 6 sources have no videos at all, 6 sources have few number of videos (respectively: 29, 7, 52, 48, 5, and 16 videos), and only 3 sources have the majority of videos (respectively: 251, 153, and 238 videos). For those which have the majority of videos, they have the same *mode* and similar *mean* and *median*. However, the three mea-

asures of these sources are close to the measures of the whole participants' evaluation for videos retrieved using user tags.

It is notable that the contribution of the sources in the retrieved videos is consistent with their contribution in the whole data sample on which the experiment was conducted, the contribution of the sources in the data sample is shown in Figure A.10 in Appendix A.2.

### Language-based dissection

Table 7.6 compares the participants' evaluation for videos retrieved using user tags with the *language-based dissected* evaluation for videos retrieved using system tags (in a similar way to Table 7.5).

	Number of videos	Mean	Median	Mode
Participants' evaluation for videos retrieved using <b>User Tags</b>	704	3.02	2	1
EN_EN	357	2.87	1	1
IT_IT	49	3.43	3	1
EN_IT	52	3.33	3.5	1
IT_EN	16	2.69	1	1
CLUSTERING_TAGS	238	2.85	2	1

Table 7.6: The descriptive measures for the *language-based dissected* data set.

### Relation-based dissection

Table 7.7 compares the participants' evaluation for videos retrieved using user tags with the *relation-based dissected* evaluation for videos retrieved using system tags (in a similar way to Table 7.5 and Table 7.6).

	Number of videos	Mean	Median	Mode
Participants' evaluation for videos retrieved using <b>User Tags</b>	704	3.02	2	1
SYNONYMY	283	2.99	2	1
HYPERNYMY	158	2.80	1	1
SIMILAR	29	2.83	2	1
RELATED	7	3.43	3	1
TRANSLATION	67	3.21	3	1
CLUSTERING_TAGS	238	2.85	2	1

Table 7.7: The descriptive measures for the *relation-based dissected* data set.

It is noteworthy in Table 7.6 and Table 7.7 that the values of the three measures (*mean*, *median*, and *mode*) are close to each other for the major sources. At the same time, they are close to the measures of the whole participants' evaluation for videos retrieved using user tags.

Finally, in the previous three tables, the summation of the videos retrieved from the different sources is always higher than the real number of videos retrieved using system tags in general (687). This is due to the fact that the same video might be retrieved because of system tags that came from two or more sources. Therefore, the same video might be counted twice, three times, or even more.

The previous descriptive statistics were just to give an idea about the collected data in a concise and summarised way. But no inferences nor conclusions can be drawn from descriptive statistics. Furthermore, they cannot support or reject research hypotheses. Inferences can be extracted from *inferential statistics*.

## 7.4 Inferential statistics

The other aspect of statistics, termed *inferential*, allows researchers to make a generalisation about the characteristics of a population, from which a sample was drawn, based on information obtained from that sample. In other words, it is the science of inferring valid conclusions about the population using the descriptive statistics. Therefore, inferential statistics are used to answer research hypotheses [12].

In order to answer (test) research hypotheses, there is variety of statistical testing procedures that can be used. Based on each exact situation, the correct statistical test is chosen. In our case, *Wilcoxon Signed-Rank test* was selected.

### 7.4.1 Wilcoxon Signed-Rank test

The Wilcoxon Signed-Rank test (also known as the Wilcoxon Matched Pairs Signed-Ranks test) is a statistical test that is appropriate to compare two sets of scores that come from the *same* participants in two *different* occasions (e.g. from one time to another), or the same individuals are subjected to more than one condition [12, 136, 137]. *Paired data* means that the scores in the two compared groups arise from the same subjects being measured more than once [138]. Since inferential statistics use descriptive statistics as inputs to draw valid inferences, Wilcoxon Signed-Rank test uses the *median* difference for inferential purposes [138, 139].

In this experiment, the *same participants* used the *same scale* in the two compared groups of scores. But the *conditions* of the experiment were *different*; the videos in one group were retrieved using user tags whilst the videos in the other group were retrieved using system tags. Hence, the appropriate statistical test to be used in such experiment is Wilcoxon Signed-Rank test.

The interest of researchers in the output of Wilcoxon Signed-Rank test is the *significance level*, presented as *P-Value*, which ranges from 0 to 1. Statistic texts indicate that if the *P-Value* is equal to or less than 0.05 ( $P\text{-Value} \leq 0.05$ ), then *the difference between the two scores is statistically significant* [136]. Otherwise, the two compared groups show no statistically significant difference even if there are differences in the descriptive statistics.

### **The results**

Accordingly, a Wilcoxon Signed-Rank test was conducted to evaluate whether the videos retrieved using user tags and the videos retrieved using system tags differ in terms of the relatedness to the search keywords. The outcome significance level (*P-Value*) was 0.97. The probability value is not less than or equal to 0.05, so the result is not significant. Therefore, there is no statistically significant difference in the relatedness to the search keywords between the videos retrieved using the two types of tags.

In other words, the findings revealed that the use of the system tags in the search is *as valid as* the use of the user tags; both types of tags produce results at the same level of relevance to the search terms with more coverage of semantically related results. Hence, using the aforementioned algorithm (Algorithm 5.3.1 for adding system tags in tag-based system) can improve the information retrieval in tagging systems. Specifically, it can address the problem where some related results exist but not retrieved due to the lack of comprehensive tags (metadata).

### **Collapsed data**

In order to support the results of Wilcoxon Signed-Rank test on the whole data set, another Wilcoxon Signed-Rank test was conducted on the collapsed version of data. The outcome significance level (*P-Value*) was 0.62. The probability value is not less than or equal to 0.05, so the result is not significant. Therefore, even after collapsing the data, the same



results were found; there is no statistically significant difference in the relatedness to the search keywords between the videos retrieved using the two types of tags.

### **Dissected data**

Furthermore, a Wilcoxon Signed-Rank test was conducted to test whether there is a significant statistical difference between the dissected participants' evaluation for the videos retrieved using system tags (prepared in Section 7.2.2) and the participants' evaluation for the videos retrieved using user tags. The inferential statistics was calculated for the dissected data in a similar way of calculating the descriptive statistics presented in Section 7.3.2. That is; the participants' evaluation for videos retrieved using user tags was compared, using Wilcoxon Signed-Rank test, with the *source-based*, the *language-based*, and the *relation-based* dissected evaluation for videos retrieved using system tags.

### **Source-based dissection**

The outcome significance levels (P-Value) of comparing the whole participants' evaluation with each *subset* of *source-based* dissected data, using Wilcoxon Signed-Rank test, are summarised in Table 7.8.

	Comparing each <i>subset</i> with the participants' evaluation for videos retrieved using user tags
EN_EN_SYNONYMS	<i>P-Value</i> = 0.64
EN_EN_HYPERNYM	<i>P-Value</i> = 0.13
EN_EN_SIMILAR	<i>P-Value</i> = 0.42
EN_EN_RELATED	<i>P-Value</i> = 0.13
EN_IT_TRANSLATION	<i>P-Value</i> = 0.26
EN_IT_SIMILAR	N/A
EN_IT_RELATED	N/A
IT_IT_SYNONYMS	<i>P-Value</i> = 0.12
IT_IT_HYPERNYM	<i>P-Value</i> = 0.26
IT_IT_SIMILAR	N/A
IT_IT_RELATED	N/A
IT_EN_TRANSLATION	<i>P-Value</i> = 0.44
IT_EN_SIMILAR	N/A
IT_EN_RELATED	N/A
CLUSTERING_TAGS	<i>P-Value</i> = 0.34

Table 7.8: The *P-Value(s)* of comparing each *subset* in the *source-based* dissected data with the participants' evaluation for videos retrieved using user tags.

**Language-based dissection**

The outcome significance levels (P-Value) of comparing the whole participants' evaluation with each *subset* of *language-based* dissected data, using Wilcoxon Signed-Rank test, are summarised in Table 7.9.

	Comparing each <i>subset</i> with the participants' evaluation for videos retrieved using user tags
EN_EN	<i>P-Value</i> = 0.58
IT_IT	<i>P-Value</i> = 0.44
EN_IT	<i>P-Value</i> = 0.89
IT_EN	<i>P-Value</i> = 0.53
CLUSTERING_TAGS	<i>P-Value</i> = 0.34

Table 7.9: The *P-Value(s)* of comparing each *subset* in the *language-based* dissected data with the participants' evaluation for videos retrieved using user tags.

**Relation-based dissection**

The outcome significance levels (P-Value) of comparing the whole participants' evaluation with each *subset* of *relation-based* dissected data, using Wilcoxon Signed-Rank test, are summarised in Table 7.10.

	Comparing each <i>subset</i> with the participants' evaluation for videos retrieved using user tags
SYNONYMY	<i>P-Value</i> = 0.47
HYPERNYMY	<i>P-Value</i> = 0.52
SIMILAR	<i>P-Value</i> = 0.74
RELATED	<i>P-Value</i> = 0.26
TRANSLATION	<i>P-Value</i> = 0.45
CLUSTERING_TAGS	<i>P-Value</i> = 0.34

Table 7.10: The *P-Value(s)* of comparing each *subset* in the *relation-based* dissected data with the participants' evaluation for videos retrieved using user tags.

As seen in the previous tables, the probability value for the dissected subsets in the

three groups is not less than or equal to 0.05, so the results are not significant. Therefore, even after dissecting the data, the same results were found; there is no statistically significant difference in the relatedness to the search keywords between the videos retrieved using user tags and the videos retrieved using dissected system tags.

## 7.5 The findings and our research hypotheses

Back to our research hypotheses in chapter one, the first hypothesis is:

H1: Adding system tags as metadata can retrieve results that are related to the searching keywords when searching in tag-based systems

This hypothesis is supported (by rejecting its null hypothesis). That is; system tags could retrieve related results. Figure B.3 shows that 41% of the participants' evaluation for videos retrieved using system tags was 4, 5, or 6. The low percentage of related videos (less than 50%) is not due to bad quality of system tags. The real reason behind this percentage is, as aforesaid, that the restriction on the searchable metadata we made (only tags), and the experiment was conducted on a subset of the videos published on YouTube website. Therefore, it was expected that several search terms will not find matching videos. 41% as abstract percentage is low, but when it is compared to the equivalent percentage for the user tags, we find that 40% of the participants' evaluation for videos retrieved using user tags was 4, 5, or 6 as seen in Figure B.2. Obviously, the reason of this percentage is not the system tags themselves as discussed.

The second hypothesis is:

H1<sub>0</sub>: The degree of relatedness between the results retrieved using system tags and the search keywords is **the same as** or **higher than** the degree of relatedness between the results retrieved using user tags and the search keywords

This hypothesis is supported (by rejecting its null hypothesis). That is; the degree of relatedness between the results retrieved using system tags and the search keywords was *the same* as the degree of relatedness between the results retrieved using user tags and the search keywords. *The same* here means that the inferential statistical test could not detect a significant difference between the two groups of results.

## 7.6 Summary

After collecting the data, using the online environment, and storing it in MySQL database, different tools were used to import the data from the database and getting it ready for analysis. Preparing the data for analysis included some data manipulation; namely, data collapsing and data dissection.

Collapsing the data is a common technique used in statistics to produce dichotomous data which facilitates the data analysis. Dissecting the data gave this research an anatomic dimension for more intelligibility and understanding of the system tags sources. Three criteria were set for dissecting the data; source-based, language-based, and relation-based.

For the whole original data set, the collapsed version of the data, and the dissected version of the data, two aspects of statistics were discussed; descriptive statistics and inferential statistics. The former statistics describe, in summarised fashion, the sample data where the latter can infer conclusions to be generalised from the sample to the population.

Using Wilcoxon Signed-Rank statistical test, the participants' evaluation of the videos retrieved using user tags and their evaluation of the videos retrieved using system tags were compared (including the collapsed and dissected data). The conclusion drawn from the sample indicates that the system tags, proposed in this work, are as valid as the user

tags and can improve the information retrieval in tag-based systems with more coverage of semantically related results. That is; system tags helps in solving the problem where relevant content are exist but not retrieved due to the insufficiency of annotating meta-data.

## Chapter 8

# Conclusion and Future Work

### *Objectives:*

---

- Providing a summary of our research.
  - Highlighting the original contributions to knowledge.
  - Comparing our work with existing related work.
  - Presenting the potential future work beyond this thesis.
-

## 8.1 Research summary

To produce this thesis, we started by giving an overview of the areas of this research; namely, Social Web and Semantic Web. Since this work is investigating the use of lexical ontologies to help addressing some shortcomings in the social tagging systems, the main concepts about the ontologies were presented. The lexical ontologies WordNet and MultiWordNet were explored in detail. Critical discussion about the social tagging was provided with emphasis on its strengths and weaknesses. However, the trade-offs between Social Web and Semantic Web were discussed (Chapter 2).

The related work in tagging area was built based on well-known characteristics of tagging system and taggers behaviours. Therefore, it was important to present various studies that address the main characteristics, features, and pattern of the social tagging. These studies give statistical information that is required to understand the related work in this area of knowledge. Afterwards, the related work was classified into three main approaches; ontological approach, social networks approach, and visualisation approach. The criterion used for this classification is the tools, or technologies, used for addressing tagging drawbacks (Chapter 3).

Having reviewed the literature in our research area, we suggested a set of rules to be followed for the successful addressing of tagging obstacles. Afterwards, we built a generic architecture for tagging systems. The architecture emphasises on addressing the challenges of social tagging with respect to the criteria (rules) we suggested. The architecture presents new semantic features in assistance with lexical semantic ontologies. Furthermore, it uses the power of the Semantic Web to introduce a solution for the multilinguality problem. In addition, our architecture can address the emergent problem of shorthand writing usage in the social tagging communities (Chapter 4).



The main components of our generic architecture were implemented in a prototype system. In the prototype system, we built our own database to host videos that were imported from YouTube. The user tags associated with these videos were also imported and stored in the database. For each user tag, our algorithm adds a number of system tags that came from either semantic ontologies (WordNet or MultiWordNet), or from tag clusters that are imported from Flickr website. Therefore, each system tag added to annotate the imported videos has a relationship with one of the user tags on that video. The relationship might be one of the following: synonymy, hypernymy, similar term, related term, translation, or clustering relation. Pseudocode algorithms are provided for our algorithm and its sub-procedures (Chapter 5).

The database, which contains videos annotated by real user tags and our proposed system tags, was exposed to end users in order to search, retrieve, and evaluate videos. This was achieved through a Web environment that we developed for the purpose of this research. The purpose of this experiment is to test the validity of our algorithm in adding system tags, or more likely, to test the added system tags themselves. By testing system tags, we investigate whether they can be considered as metadata in which users can search and retrieve related results. Relatedness of results is a relative and subjective criterion. Therefore, the relatedness of the videos retrieved using system tags should be compared with a relatedness of videos that are retrieved using another kind of metadata. In this experiment, user tags were considered to be the other kind of metadata for comparison purposes. However, for the comparison to be fair, the same searching algorithm, the same subjects, the same search keywords, and the same evaluation metrics (Likert scale) were considered (Chapter 6).

The users' evaluation of the two groups of retrieved videos is collected and saved in our database. The collected data was prepared and exported to the SPSS to be analysed.

The data analysis produced two types of statistics; descriptive statistics and inferential statistics. The former describes the collected sample of data, while the latter infer conclusions to be generalised from the sample to the population. The inferential part was done using a well-known statistical test called *Wilcoxon Signed-Rank*. The results revealed that there is no statistically significant difference in term of relatedness between the two groups of videos. In other words; both user tags and system tags can retrieve videos that are at the same level of relatedness to the user search terms. Therefore, system tags can be used to address the problem where related contents exist in the tagging system but they are not being retrieved due to the lack of semantic annotation. By dissecting the system tags resources, we could not detect any statistically significant variation among them in term of the system tags quality (Chapter 7).

## 8.2 Success criteria revisited

As mentioned in Section 1.5, supporting or rejecting our research hypotheses verifies whether the system tags can improve the information retrieval in tagging systems or not. As presented in Section 7.5, both research hypotheses were supported; new system tags are as valid as user tags with more coverage of semantically related results. Therefore, in tag-based systems, system tags can be considered a successful solution for addressing the challenges of semantic relations, multilinguality, and shorthand written tags.

## 8.3 Contribution to knowledge

The main contributions to knowledge in this work are summarised as follows:

1. **Generic architecture for tagging systems:** We built a generic architecture for tagging systems. This architecture can be considered as a template for building any tagging system. Our architecture can address the majority of tagging systems

drawbacks.

2. **Set of criteria:** Having reviewed the extensive literature in the area of tagging systems and Social Web, we could formulate rules, or standards, for any approach that tries to address the challenges in tagging systems. If followed, these rules can keep the integrity and ethos of tagging systems.
3. **Addressing the problem of semantic relations in tagging systems:** One of the main problems in tagging systems is the lack of semantic relations (e.g. synonyms). Our approach showed promising results in addressing this problem by using the lexical ontology WordNet.
4. **Addressing the problem of multilinguality in tagging systems:** One of the main problems in tagging systems is the lack of cross-language information retrieval. Our approach showed promising results in addressing this problem by using the lexical ontology MultiWordNet.
5. **Addressing the problem of shorthand writing tags:** One of the main problems in tagging systems is the use of shorthand words (tags). Our approach showed promising results in addressing this problem by using tag clusters to define a context for such tags.

Addressing the abovementioned problems improves the information retrieval in tagging systems. Consequently, the previously invisible related content in tagging systems can now be retrieved.

### 8.4 Comparison with existing related work

One of the main strengths of our proposed solution is that we maintain the users' tags integrity. In other words, the tags that are originally provided by users will not be modified

nor deleted. In [5], they built a bottom-up ontology from folksonomy that is used for user tags auto-replacement. We argue that changing or updating users' tags is not acceptable; users will be dissatisfied if they add some tags and the next day they discover that the system is changing these tags.

Moreover, our solution keeps the interaction pattern between the taggers and the tagging system. That is; users are not hindered with unconventional way of tagging. In [7], for instance, they addressed the lack of semantics by putting more effort on the user to give a classification of tags. While in [106] the user is given suggestions and is asked to give feedback about these suggestions. The interaction pattern between the users and the Web was one of the main reasons behind the involvement of vast numbers of users in social tagging. Changing this pattern contradicts with the ethos of tagging; which is *providing simple free text words*.

To our knowledge, the notion of multilinguality is not addressed in the related work. Rather, they intended to remove the non-English words in [4] while trying to create semantic metadata.

Another feature of our architecture is integrating the social and semantic resources to enhance the semantics of social tagging. Related research attempts depend either on social resources (e.g. [8] and [78]) or on semantic resources (e.g. [6]).

To evaluate our solution, we exposed 29,770 user tags and 36,038 system tags to 204 real users. Whereas, for example, only 10 tags were used in [78], and no empirical results were provided in [9].

Furthermore, we take advantage of WordNet ontology without bothering the taggers

with its rigidity (e.g. like in [7] and [6]). Rather, the ontology in our architecture is used for computational processes that are executed in the system background.

## 8.5 Limitations and Future work

Several future research directions can be expanded for the work presented in this thesis. The potential future work is summarised as follows:

- Our experiment was conducted on videos imported from YouTube. Applying the same algorithm and experiment on different kind of data (e.g. Photos on Flickr, URLs on Del.icio.us) might give different results due to differences in the nature of the evaluated data.
- In our experiment, the semantic resources (WordNet and MultiWordNet) were queried by the user tags to retrieve its relevant system tags. Performing normalisation on the user tags before querying the semantic resources (e.g. stemming) might give better quality of system tags.
- The semantic multilingual ontology (MultiWordNet) used in our experiment includes only two languages; English and Italian. Using other multilingual ontologies that cover more languages (e.g. EuroWordNet) might give better results. We argue that multilinguality can provide better results if applied on photo tagging systems (e.g. Flickr). Being English and searching using an English word in a tagging system and retrieving contents that contain written or spoken Russian language might seem irrelevant. This is not the case when the retrieved content is a photo since photos do not contain lingual data.
- In our experiment, existing Flickr clusters were used. As aforementioned, the clustering criteria used to group tags together have not been released officially by Flickr. Therefore, importing huge tag set from online tagging systems (e.g. Del.icio.us)

and building the clusters from scratch might give better results. Tag co-occurrence might be considered as a clustering criterion.

- Beside the challenges addressed by our architecture, disambiguating the polysemous tags might give better results.
- Time-wise and space-wise issues are beyond the scope of this thesis. Another research direction is to investigate the effect of adding system tags in terms of the time consumed while searching and the space occupied by system tags.

# Bibliography

- [1] George Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(1):39 – 41, 1995.
- [2] Christiane Fellbaum, Derek Gross, and Katherine Miller. Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265 – 277, 1990.
- [3] Wikipedia.com. Power Law. Retrieved 20 October 2009 <[http://en.wikipedia.org/wiki/Power\\_law](http://en.wikipedia.org/wiki/Power_law)>.
- [4] Hend Al-Khalifa and Hugh Davis. FolksAnnotation: A semantic metadata tool for annotating learning resources using folksonomies and domain ontologies. In *Proceedings of the Conference on Innovations in Information Technology*, 2006.
- [5] Fawaz Ghali, Mike Sharp, and Alexandra Cristea. Folksonomies and ontologies in authoring of adaptive hypermedia. In *A3H: 6th International Workshop on Authoring of Adaptive and Adaptable Hypermedia Workshop*, 2008.
- [6] David Laniado, Davide Eynard, and Marco Colombetti. Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Semantic Web Application and Perspectives - 4th Italian Semantic Web Workshop*, volume 314 of *CEUR Workshop Proceedings*, 2007.

- [7] Sun-Sook Lee and Hwan-Seung Yong. OntoSonomy: Ontology-based extension of folksonomy. In *Proceedings of the 2008 IEEE International Workshop on Semantic Computing and Applications*, pages 27 – 32, 2008.
- [8] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the 15th International World Wide Web Conference*, volume 6, 2006.
- [9] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, pages 31 – 40, 2006.
- [10] Yusef Hassan-Montero and Victor Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of Multidisciplinary Information Sciences and Technologies*, 2006.
- [11] Paul Leedy and Jeannne Ormrod. *Practical Research: Planning and Design*. Upper Saddle River, NJ: Merrill Prentice Hall, 7th edition, 2001.
- [12] Robert Burns. *Introduction to research methods*. Sage Publications, 4th edition, 2000.
- [13] Dieter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. The MIT Press, 2005.
- [14] Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, 2008.
- [15] Tim Berners-Lee. *Weaving the Web: The Past, Present, and Future of the World Wide Web, by its Inventor*. Texere Publishing Ltd., 1999.



- [16] James Hendler, Tim Berners-Lee, and Eric Miller. Integrating applications on the Semantic Web. *Journal of the Institute of Electrical Engineers of Japan*, 122(10):676 – 680, 2002.
- [17] W3schools.com. Semantic Web. Retrieved 10 October 2009 <<http://www.w3schools.com/semweb/default.asp>>.
- [18] John Davies, Dieter Fensel, and Frank Van Harmelen. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. Wiley, 2003.
- [19] Thomas Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220, 1993.
- [20] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Twente, Enschede, 1997.
- [21] Rudi Studer, Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data and knowledge engineering*, 25:161 – 197, 1998.
- [22] George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235 – 244, 1990.
- [23] Satanjeev Banerjee and Satanjeev Banerjee. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136 – 145, 2002.
- [24] Richard Beckwith, George Miller, and Randee Tengi. Design and implementation of the WordNet lexical database and searching software. Technical report, Princeton University Cognitive Science Laboratory, 1993.

- [25] Piek Vossen. WordNet, EuroWordNet and global WordNet. *Revue Francaise de Linguistique Appliquee / RFLA*, 7(1), 2002.
- [26] Jen-Yi Lin, Chang-Hua Yang, Shu-Chuan Tseng, and Chu-Ren Huang. The structure of polysemy: A study of multi-sense words based on WordNet. In *Proceedings of the 16th Pacific Asia Conference on Language, Information, and Computation*, pages 320 – 329, 2002.
- [27] George Miller. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245 – 264, 1990.
- [28] Christiane Fellbaum. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278 – 301, 1990.
- [29] Chiao-Shan Lo, Yi-Rung Chen, Chih-Yu Lin, and Shu-Kai Hsieh. Automatic labeling of troponymy for Chinese verbs. In *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing*, 2008.
- [30] Sara Mendes. Adjectives in WordNet.PT. In *Proceedings of the Global WordNet Association Conference*, pages 225 – 230, 2006.
- [31] Sunkyoung Baek, Miyoung Cho, and PanKoo Kim. Matching colors with KANSEI vocabulary using similarity measure based on WordNet. In *International Conference on Computational Science and its Applications*, pages 37 – 45, 2005.
- [32] Kyoko Kanzaki, Francis Bond, Takayuki Kuribayashi, and Hitoshi Isahara. Enriching the adjective domain in the Japanese WordNet. In *Proceedings of the 7th International Conference on Natural Language Processing*, volume 6233, pages 162 – 166, 2010.
- [33] Satanjeev Banerjee. Adapting the Lesk algorithm for word sense disambiguation to WordNet. Master’s thesis, University of Minnesota, USA, 2002.

- [34] Jorge Morato, Miguel Marzal, Juan Llorns, and Jos Moreiro. WordNet applications. In *Proceeding of the 2nd Global WordNet Conference*, 2004.
- [35] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on Global WordNet*, 2002.
- [36] Peter Mika. *Social Networks and the Semantic Web*, volume 5 of *Semantic Web and Beyond*. Springer - Verlag, 2007.
- [37] Thomas Gruber. Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics*, 6(1):4 – 13, 2008.
- [38] Fabio Abbattista, Fabio Calefato, Domenico Gendarmi, and Filippo Lanubile. Shaping personal information spaces from collaborative tagging systems. In *Proceedings of the 3rd International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, volume 4694 of *Lecture Notes in Computer Science*, pages 728 – 735, 2007.
- [39] Scott Bateman, Christopher Brooks, and Gord Mccalla. Collaborative tagging approaches for ontological metadata in adaptive e-Learning systems. In *Proceedings of the Workshop on Applications of Semantic Web Technologies for e-Learning, at the 4th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, pages 3 – 12, 2006.
- [40] Alan Dix, Stefano Levialdi, and Alessio Malizia. Semantic Halo for collaboration tagging systems. In *Workshop on the Social Navigation and Community-Based Adaptation Technologies*, 2006.
- [41] Scott Golder and Bernardo Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198 – 208, 2006.

- [42] Hak-Lae Kim, Simon Scerri, John Breslin, Stefan Decker, and Hong-Gee Kim. The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2008.
- [43] Thomas Knerr. Tagging ontology - towards a common ontology for folksonomies. Retrieved 20 August 2009 <<http://tagont.googlecode.com/files/TagOntPaper.pdf>>.
- [44] Qingfeng Li and Stephen Lu. Collaborative tagging applications and approaches. *Multimedia*, 15(3):14 – 21, 2008.
- [45] Gregor Macgregor and Emma Mcculloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5), 2006.
- [46] Zhichen Xu, Yan Fu, Jianchang Mao, and Difu Su. Towards the Semantic Web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW*, 2006.
- [47] Thomas Vander Wal. Explaining and showing broad and narrow folksonomies. 2005.
- [48] Andrea Marchetti, Maurizio Tesconi, Francesco Ronzano, Marco Rosella, and Salvatore Minutoli. Semkey: A semantic collaborative tagging system. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organisation*, 2007.
- [49] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 522 – 536, 2005.

- [50] Marti Hearst and Daniela Rosner. Tag clouds: Data analysis tool or social signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 160, 2008.
- [51] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualisation. *Clinical Orthopaedics and Related Research*, abs/cs/0703109, 2007.
- [52] Walky Rivadeneira, Gruen Daniel, Michael Muller, and David Millen. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *Proceedings of the Special Interest Group in Human-Computer Interaction Conference on Human Factors in Computing Systems*, pages 995 – 998, 2007.
- [53] Daniel Steinbock, Roy Pea, and Byron Reeves. Wearable tag clouds: Visualisations to facilitate new collaborations. In *Proceedings of the 8th international Conference on Computer Supported Collaborative Learning*, pages 672 – 674, 2007.
- [54] Cyprien Lomas. 7 things you should know about social bookmarking. *The EDUCAUSE Learning Initiative*, 2005.
- [55] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population*, pages 39 – 43, 2008.
- [56] Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analysing social bookmarking systems: A del.icio.us cookbook. In *Proceedings of Mining Social Data Workshop*, pages 26 – 30, 2008.
- [57] Adam Mathes. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication - LIS590CMC*, 2004.

- [58] Valentina Malaxa and Ian Douglas. A framework for metadata creation tools. *Interdisciplinary Journal of Knowledge and Learning Objects*, 1:151 – 162, 2005.
- [59] Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart Weibel. Metadata principles and practicalities. *D-lib Magazine*, 8(4):1 – 16, 2002.
- [60] Gary Geisler, Sarah Giersch, David McArthur, and Marilyn McClelland. Creating virtual collections in digital libraries: Benefits and implementation issues. In *Proceedings of the Joint Conference on Digital Libraries*, pages 210 – 218, 2002.
- [61] Lyndsay Greer. The learning matrix: Cataloging resources with rich metadata. In *Proceedings of the 2nd Joint Conference on Digital Libraries*, page 375, 2002.
- [62] Carol Hert. Studies of metadata creation and usage. Technical report, School of Information Studies, Syracuse University, 2001.
- [63] Catherine Marshall. Making metadata: A study of metadata creation for a mixed physical-digital collection. In *Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 162 – 171, 1998.
- [64] Sarah Hayman. Folksonomies and tagging: New developments in social bookmarking. In *Proceedings of the Ark Group Conference: Developing and Improving Classification Schemes*, 2007.
- [65] Saba Anila. Collaborative tagging: A new way of defining keywords to access Web resources. In *Proceedings of International Convention on Automation of Libraries in Education and Research: From Automation to Transformation*, pages 309 – 315, 2008.
- [66] Eetu Makela. Harnessing folksonomies for search. In *Proceedings of the Seminar on Web 2.0*, 2006.

- [67] Alexander Kreiser, Andreas Nauerz, Fedor Bakalov, Birgitta Konig-Ries, and Martin Welsch. A Web 3.0 approach for improving tagging systems. In *Proceedings of the International Workshop on Web 3.0: Merging Semantic Web and Social Web (in Conjunction with the 20th International Conference on Hypertext and Hypermedia)*, 2009.
- [68] Sihem Amer-Yahia, Michael Benedikt, and Philip Bohannon. Challenges in searching online communities. *IEEE Data Engineering Bulletin*, 30(2):23 – 31, 2007.
- [69] Ikki Ohmukai, Masahiro Hamasaki, and Hideaki Takeda. A proposal of community-based folksonomy with RDF metadata. In *Proceedings of the Workshop on End User Semantic Web Interaction, held in conjunction with the International Semantic Web Conference*, 2005.
- [70] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalised recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259 – 266, 2008.
- [71] Elizeu Santos-Neto, Matei Ripeanu, and Adriana Iamnitchi. Tracking usage in collaborative tagging communities. In *Proceedings of the Workshop on Contextualised Attention Metadata*, 2007.
- [72] Sofia Angeletou, Marta Sabou, and Enrico Motta. Semantically enriching folksonomies with FLOR. In *Proceedings of the European Semantic Web Conference Workshop*, 2008.
- [73] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*, pages 211 – 220, 2007.

- [74] Siegfried Handschuh, Steffen Staab, and Rudi Studer. Leveraging metadata creation for the Semantic Web with CREAM. In *Proceedings of KI 2003: Advances in Artificial Intelligence: 26th Annual German Conference on AI*, volume 2821 of *Lecture Notes in Computer Science*, pages 19 – 33, 2003.
- [75] Jeff Pan, Stuart Taylor, and Edward Thomas. Reducing ambiguity in tagging systems with folksonomy search expansion. In *Proceedings of the 6th Annual European Semantic Web Conference*, pages 669 – 683, 2009.
- [76] Valentin Robu, Harry Halpin, and Hana Shepherd. Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, 3(4), 2009.
- [77] Kilian Weinberger, Malcolm Slaney, and Roelof Zwol. Resolving tag ambiguity. In *ACM Multimedia*, pages 111 – 120, 2008.
- [78] Ching Yeung, Nicholas Gibbins, and Nigel Shadbolt. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pages 251 – 260, 2009.
- [79] Marieke Guy and Emma Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1), 2006.
- [80] Christopher Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th International Conference on World Wide Web*, pages 625 – 632, 2006.
- [81] Fernando Sanchez-Zamora and Martin Llamas-Nistal. Visualising tags as a network of relatedness. In *Proceedings of the 39th ASEE/IEEE Frontiers in Education Conference*, 2009.
- [82] Clay Shirky. *Ontology is overrated: Categories, links, and tags*. 2005.



- [83] Rachel Or-Bach. Collaborative tagging, metadata creation and learning - a study within a higher-education course. *Chais Research Center for the Integration of Technology in Education, The Open University of Israel*, 2007.
- [84] Peter Merholz. Clay Shirky's Viewpoints are Overrated. *PETERME.COM*. Retrieved 20 October 2009 <<http://www.peterme.com/archives/000558.html>>.
- [85] David Weinberger. (2006). PennTags - When Card Catalogs Meet Tags. *Many2Many: a Group Weblog on Social Software*. Retrieved 20 October 2009 <[http://many.corante.com/archives/2006/06/10/penntags\\_when\\_card\\_catalogs\\_meet\\_tags.php](http://many.corante.com/archives/2006/06/10/penntags_when_card_catalogs_meet_tags.php)>.
- [86] Stuart Weibel. (2006). Hybrid Vigor. *Weibel Lines*. Retrieved 20 October 2009 <[http://weibel-lines.typepad.com/weibelines/2006/03/hybrid\\_vigor.html](http://weibel-lines.typepad.com/weibelines/2006/03/hybrid_vigor.html)>.
- [87] Celine Van Damme, Martin Hepp, and Katharina Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Proceedings of the Bridging the Gap between Semantic Web and Web 2.0*, pages 57 – 70, 2007.
- [88] Thomas Gruber. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, 3(2):1 – 11, 2007.
- [89] Krystyna Matusiak. Towards user-centered indexing in digital image collections. *OCLC Systems & Services*, 22(4):283 – 298, 2006.
- [90] Atefeh Sharif. Combining ontology and folksonomy: An integrated approach to knowledge representation. In *Proceedings of the Emerging trends in technology: Libraries Between Web 2.0, Semantic Web, and search technology*, 2009.
- [91] Noriko Tomuro and Andriy Shepitsen. Construction of disambiguated folksonomy ontologies using wikipedia. In *Proceedings of the Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 42 – 50, 2009.

- [92] Alexander Mikroyannidis. Toward a social Semantic Web. *Computer*, 40(11):113 – 115, 2007.
- [93] Uldis Bojars, John Breslin, Vassilios Peristeras, Giovanni Tummarello, and Stefan Decker. Interlinking the Social Web with semantics. *IEEE Intelligent Systems*, 23(3):29 – 40, 2008.
- [94] Uldis Bojars, Alexandre Passant, John Breslin, and Stefan Decker. Data portability with SIOC and FOAF. In *XTech*, 2008.
- [95] Margaret Kipp and Grant Campbell. Patterns and inconsistencies in collaborative tagging systems : An examination of tagging practices. 2006.
- [96] Harry Halpin, Valentin Robu, and Hana Shepard. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop*, 2006.
- [97] Thomas Gruber. TagOntology - a way to agree on the semantics of tagging data. Retrieved 20 August 2009 <<http://tomgruber.org/writing/tagontology.htm>>.
- [98] Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on “Network Analysis in Natural Sciences and Engineering”*, 20(4):245 – 262, 2007.
- [99] Nitin Borwankar. (2006). *Slicing and dicing data 2.0: Foundation data dodel for folksonomy navigation*. Retrieved 09 November 2009 <<http://tagschema.com/blogs/tagschema/2005/06/slicing-and-dicing-data-20-part-2.html>>.
- [100] Henry Story. (2007). *Search, Tagging, and Wikis*. Retrieved 09 November 2009 <[http://blogs.oracle.com/bblfish/entry/search\\_tagging\\_and\\_wikis](http://blogs.oracle.com/bblfish/entry/search_tagging_and_wikis)>.

- [101] Rcharad Newman, Danny Ayers, and Seth Russell. (2005). *Tag Ontology*. Retrieved 09 November 2009 <[www.holygoat.co.uk/owl/redwood/0.1/tags/](http://www.holygoat.co.uk/owl/redwood/0.1/tags/)>.
- [102] Alexandre Passant and Philippe Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW Workshop Linked Data on the Web*, 2008.
- [103] Simon Scerri, Michael Sintek, Ludger van Elst, and Siegfried Handschuh. (2007). *NEPOMUK Annotation Ontology*. Retrieved 09 November 2009 <<http://www.semanticdesktop.org/ontologies/nao/>>.
- [104] Hak Lae Kim, Alexandre Passant, John Breslin, Simon Scerri, and Stefan Decker. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Proceedings of the 2nd International Conference on Semantic Computing*, pages 315 – 322, 2008.
- [105] Silvia Bindelli, Claudio Criscione, Carlo Curino, Mauro Drago, Davide Eynard, and Giorgio Orsi. Improving search and navigation by combining ontologies and social tags. In *Proceedings of the 1st International Workshop on Ambient Data Integration*, 2008.
- [106] Kees Sluijs and Geert-Jan Houben. Relating user tags to ontological information. In *Proceedings of 5th International Workshop on Ubiquitous User Modeling*, 2008.
- [107] Glenn Fung. A comprehensive overview of basic clustering algorithms. Technical report, Department of Computer Sciences, University of Wisconsin-Madison, 2001.
- [108] Terrell Russell. Cloudalicious: Folksonomy over time. In *Proceedings of the 6th Joint Conference on Digital Libraries*, pages 364 – 364, 2006.

- [109] The Lancaster Stemming Algorithm. Retrieved 20 June 2011 <<http://www.comp.lancs.ac.uk/computing/research/stemming/index.htm>>.
- [110] Flickr Customer Care, case1659812@support.flickr.com, 2010. Re: [Flickr Case 1659812] Re: Other issues. [email] Message to M. Magableh (murad415@yahoo.com). Sent Tuesday 16 November 2010.
- [111] KARL. (2005). *Leveraging folksonomy - Flickr clusters*. Retrieved 16 November 2010 <<http://blog.experiencecurve.com/archives/leveraging-folksonomy-flickr-clusters>>.
- [112] Princeton University. Retrieved 2 February 2011 <<http://wordnet.princeton.edu/man/grind.1WN.html>>.
- [113] The EDUCAUSE Learning Initiative. 7 things you should know about YouTube. 2009.
- [114] Thomas Connolly, Carolyn Begg, and Anne Strachan. *Database Systems: A Practical Approach to Design, Implementation and Management*. Addison-Wesley, 2nd edition, 1998.
- [115] Floyd Fowler. *Survey research methods*. Applied social research methods series. Sage Publications, 2009.
- [116] Joseph Abramson and Zvi Abramson. *Survey Methods in Community Medicine: Epidemiological Research, Programme Evaluation, Clinical Trials*. Churchill Livingstone, 5th edition, 1999.
- [117] Pamela Alreck and Robert Settle. *The Survey Research Handbook*. McGraw Hill, 2nd edition, 1995.
- [118] Research Methods Knowledge Base. Likert Scaling. Retrieved 29 March 2011 <<http://www.socialresearchmethods.net/kb/scallik.php>>.

- [119] Carolyn Preston and Andrew Colman. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1–15, 2000.
- [120] David Freedman. Sampling. *Encyclopedia of Social Science Research Methods*, 3:986 – 990, 2004.
- [121] David De Vaus. *Surveys in social research*. UCL Press, 1991.
- [122] Nancy Burns and Susan Grove. *Understanding Nursing Research*. WB Saunders Co, 2nd edition, 1999.
- [123] Experiment-Resources.com. Convenient sampling applied to research. Retrieved 15 April 2011 <<http://www.experiment-resources.com/convenience-sampling.html>>.
- [124] Sampling. The How-To's of Monitoring and Evaluation. Retrieved 15 April 2011 <<http://www.fhi.org/NR/rdonlyres/etdgabwszyyk2hmkqosvl2mieeatan6rrj4l4lfuv52dlbt7knrewo6qfzosuzq7raxy63chxkz32c/chapter6.pdf>>.
- [125] Non-Probability Samples. Retrieved 15 April 2011 <<http://www.tardis.ed.ac.uk/~kate/qmcweb/s8.htm>>.
- [126] Denise Polit and Bernadette Hungler. *Nursing Research Principles and Methods*. Lippincott, 6th edition, 1999.
- [127] Harvey Motulsky. *The InStat Guide to Choosing and Interpreting Statistical Tests*. GraphPad Software, Inc, 2001.
- [128] Franz Faul and Edgar Erdfelder (1992) GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS (computer programme). Bonn, FRG: Bonn University, Department of Psychology.

- [129] Edwin Teijlingen and Vanora Hundley. The importance of pilot studies. *Nurs Stand*, 16(40):33 – 36, 2002.
- [130] Michael Allen, Tricia Coulter, Carol Dwyer, Laura Goe, John Immerwahr, Amy Jackson, Jean Jonson, Regina Oliver, Amber Ott, Daniel Rischly, Jonathan Rochkind Cortney Rowland, and Susan Smartt. America’s challenge: Effective teachers for at-risk schools and students. Technical report, National Comprehensive Center for Teacher Quality, 2007.
- [131] Rebecca Snyder, James Bills, Sharon Phillips, Margaret Tarpley, and John Tarpley. Specific interventions to increase women’s interest in surgery. *Journal of the American College of Surgeons*, 207(6):942 – 947, 2008.
- [132] Polling the Nations. Intensity of feeling. Retrieved 9 June 2011 <<http://poll.orspub.com/static.php?type=about&page=questions6>>.
- [133] Peter Grimbeek, Fiona Bryer, Wendi Beamish, and Michelle D’Netto. Use of data collapsing strategies to identify latent variables in questionnaire data: Strategic management of junior and middle school data on the CHP questionnaire. In *Proceedings of the 3rd Annual International Conference on Cognition, Language and Special Education*, pages 1 – 15, 2005.
- [134] Laerd Statistics. Descriptive and Inferential Statistics. Retrieved 13 June 2011 <<http://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php>>.
- [135] Descriptive and Inferential Statistics: Summary. Retrieved 13 June 2011 <<http://www.habermas.org/stat2f98.htm>>.
- [136] Julie Pallent. *SPSS Survival Manual*. Open University Press, 2nd edition, 2005.

## BIBLIOGRAPHY

---

- [137] Laerd Statistics. Wilcoxon Signed Rank Test using SPSS. Retrieved 19 June 2011 <<http://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php>>.
- [138] Information Point. Wilcoxon Signed Rank Test. Retrieved 19 June 2011 <[http://www.blackwellpublishing.com/specialarticles/jcn\\_9\\_584.pdf](http://www.blackwellpublishing.com/specialarticles/jcn_9_584.pdf)>.
- [139] Wilcoxon Signed-Ranks Test for the Median Difference. Retrieved 19 June 2011 <[http://courses.wcupa.edu/rbove/Berenson/CD-ROM%20Topics/topice-10\\_5.pdf](http://courses.wcupa.edu/rbove/Berenson/CD-ROM%20Topics/topice-10_5.pdf)>.

# Appendix A

## The Experiment Sample Data

### A.1 The keywords used to import the YouTube videos

```
// ----- //
```

```
                ** The English Keywords **
```

```
// ----- //
```

```
"education", "tutorial", "research", "student", "academy", "learning",  
"technology", "system", "computer", "computing", "programming", "web",  
"internet", "software", "engineering", "science", "media", "video", "tv",  
"show", "music", "audio", "news", "cinema", "movie", "radio", "photo",  
"ad", "advertisement", "entertainment", "comedy", "style", "model", "art",  
"design", "beautiful", "paint", "beauty", "transportation", "car", "plane",  
"train", "flight", "travel", "tourism", "holiday", "human", "people", "man",  
"girl", "kid", "baby", "creature", "children", "arab", "social", "culture",  
"religion", "history", "dancing", "sport", "football", "games", "business",  
"product", "company", "money", "economy", "office", "mobile", "language",  
"nature", "animal", "bird", "fish", "mammal", "jungle", "life", "world",  
"health", "hospital", "military", "accommodation", "law", "utility", "event",  
"funny", "sad", "communication", "food", "drink", "dish", "restaurant",  
"beverage", "sex", "morocco", "algeria", "tunisia", "libya", "egypt", "iraq",  
"jordan", "syria", "lebanon", "palestine", "saudi arabia", "sudan", "qatar",  
"kuwait", "united arab emirates", "bahrain", "oman", "yemen", "leicester",  
"united kingdom", "london", "united states", "revolution", "protest",  
"jesus", "robot", "demonstration", "war", "kill", "alqaddafi", "obama",  
"hosni mubarak", "speech", "haifa", "christmass", "christian", "hollywood",  
"arab songs", "arab celebrity", "flamenco", "salsa", "belly dance", "theory",  
"hip hop", "matlab", "family guy", "latest movies", "facebook", "smart home",  
"honda", "iphone", "ipad", "bruce lee", "action movies", "research methods",  
"hybrid technology", "nano technology", "hybrid cars", "elearning", "diet",  
"egovernment", "hot topic", "academic", "regime", "nursing", "glass", "word",  
"twitter", "engineering", "world celebrity", "excel", "access", "windows",  
"apple", "tagging", "drawing", "statistics", "spss", "lady gaga", "avatar",  
"black swan", "james cameron", "rihanna", "beyonce", "park", "ronaldo", "OS",
```



## APPENDIX A. THE EXPERIMENT SAMPLE DATA

---

"messi", "mercedes", "audi", "toyota", "bmw", "volkswagen", "volvo", "ebay", "amazon", "youtube", "how to", "definition", "steve jobs", "tablet", "vlog", "blog", "talent", "arabs got talent", "poet", "xfactor", "big brother", "pop", "rock", "super star", "millionaire", "circus", "viol", "violin", "guitar", "disney park", "walmart", "traffic light", "prank", "boyfriend", "girl friend", "animal sex", "animal love", "heart", "blood", "kidney", "surgery", "diabetes", "cancer", "disease", "dog", "cat", "lion", "tiger", "mouse", "rabbit", "mouse", "cartoon", "animation", "simulation", "manga", "prey", "qualitative", "quantitative", "middle east", "prey", "predator", "crocodile", "giraffe", "dove", "bbc", "cnn", "aljazeera", "lol", "tomato", "potato", "junk food", "market", "take away", "fat", "slim", "trick", "magic", "cucumber", "melon", "fruit", "vegetable", "tai food", "sushi", "arabic food", "banana", "pepper", "spicy", "hot", "onion", "onion", "garlic", "indian food", "tool", "technical", "java", "c++", "OOP", "aid", "object", "TED", "study", "toefl", "ielts", "truth", "touch", "techno", "savvy", "conference", "journal", "publisher", "kebab", "toast", "cuisine", "teacher", "student", "university", "college", "help", "spare parts", "porn", "recognition", "chapter", "verse", "anthem", "national", "international", "multinational", "army", "soldier", "jacket", "t-shirt", "trouser", "outlet", "outfit", "electronics", "electricity", "sex education", "gay", "lesbian", "passion", "toys", "positions", "intimacy", "patriot", "extremist", "king", "queen", "prince", "princess", "duke", "duchess", "lawer", "band", "gang", "bond", "symphony", "yanni", "pants", "botox", "makeup", "tie", "demolition", "damage", "destroy", "hate", "anger", "angry", "lovely", "ugly", "big", "small", "tall", "short", "wet", "dry", "photoshop", "adobe", "paint", "sea", "river", "tree", "greenery", "building", "construction", "framework", "architecture", "modelling", "algorithm", "thesis", "grounded theory", "akon", "american idol", "dirty bit", "eminem", "firework", "hello", "hold", "first time", "expert", "profession", "professional", "genius", "stupid", "fool", "banned", "bastard", "condom", "horse", "horse riding", "invitation", "doctor", "laser", "hair removal", "skin care", "case study", "pilot study", "ecommerce", "ebusiness", "introduction to", "how to learn", "criminal law", "criminal", "police", "report", "survey", "questionnaire", "focus group", "click", "social community", "old people", "young people", "prostitute", "youth", "labour", "methodology", "paper", "brazil", "north america", "dubai", "paris", "new york", "washington", "fifa", "world cup", "birthday", "party", "anniversary", "trend", "english series", "english movies", "god", "islam", "population", "running", "walking", "sleep", "bed", "home", "house", "garden", "kitchen", "bathroom", "sauna", "jacuzzi", "swimming pool", "shower", "sun", "moon", "night", "day", "time", "minute", "second", "rent", "sell", "buy", "easy", "income", "input", "output", "grid computing", "network", "software engineering", "english literature", "linguistics", "mistake", "error", "user generated", "how to fix", "how to do", "database", "field", "normalisation", "SQL", "RAM", "ROM", "laptop", "HD TV", "camera", "HD digital camera", "3D TV", "virtual reality", "mattress", "viva", "PHD", "visual", "background", "image processing", "photo", "mechatronics", "yahoo", "google", "search engine", "book", "PDF", "convertor", "converter", "yummy", "delicious", "folksonomy", "sky", "cloud", "plan", "airlines", "flight", "aircraft", "host", "hostess", "first class", "engine", "helicopter", "van", "dynamic", "static", "solar", "agriculture", "terrorism", "tourist", "flag", "ministry", "prime minister", "letter", "translator", "dictionary", "city", "mountain", "valley", "hill", "hell", "paradise", "heaven", "jew", "jewish", "white", "black", "red", "yellow", "cotton", "oil", "petrol", "politics", "dentist", "hospital", "infirmity", "train", "bus", "boat", "jet ski",

"fail", "crash", "funny kid", "funny man", "naked and funny", "funny pool",  
 "funny accident", "car accident", "highway", "shouting", "crazy", "idiot",  
 "cctv", "drunk", "weekend", "holiday", "money market", "travel agency",  
 "tanning", "halal", "teasing", "annoying", "orange", "broadband", "virgin",  
 "mobile offers", "latest offers", "sale", "computer cookies", "sweet cookie",  
 "kiss", "michael jackson", "tea", "milk", "coffee", "cafe", "rotana cinema",  
 "rotana music", "rotana zaman", "mbc", "lbc", "mtv channel", "melody aflam",  
 "melody music", "art channels", "Jerusalem", "ad hoc network", "weather",  
 "mobile operator", "tv news", "newspaper", "roma", "shisha", "smoking",  
 "crying", "transplantation", "flower", "rose", "plant", "planet", "fake it",  
 "fake wedding", "mr bean", "pray", "play", "season", "cooking", "receipt",  
 "stand up comedy", "opera", "sensors", "carpet", "pet", "amazing", "huge",  
 "extremely", "water", "spring", "summer", "winter", "autumn", "leaf", "pig",  
 "boring", "keyboard", "monitor", "best deals", "self learning",  
 "remote access", "chatting", "video chatting", "msn", "skype", "voip",  
 "royal", "team", "barcelona", "amsterdam", "big capitals", "madrid", "cairo",  
 "tripoli", "answers", "question", "lifestyle", "volleyball", "tennis", "GPS",  
 "tomtom", "navigation system", "dorm", "scholarship", "grant", "scholars",  
 "invention", "innovation", "creative", "creativity", "simplicity", "folk",  
 "modern", "ancient", "patient", "patience", "sick", "ill", "nurse", "salary",  
 "pension", "retired", "hire", "fire", "benefit", "employee", "employment",  
 "manager", "management", "boss", "firm", "sme", "sms", "text messages",  
 "bag", "luggage", "leather", "heather", "discussion", "debate", "focus",  
 "spot light", "week harvest", "harvest", "green house effect", "gas",  
 "power", "muscles", "body building", "belly exercise", "mathematics",  
 "math", "physics", "physical contact", "physical effort", "eye contact",  
 "physical equation", "skills", "social skills", "plan b", "planning",  
 "decision support", "knowledge", "knowledge management", "data processing",  
 "data collection", "information", "information technology", "browsing",  
 "stone", "rock", "hat", "cover", "head", "snake", "scorpion", "fighting",  
 "dirty", "prison", "jeal", "investigation", "proof", "CEO", "chair", "table",  
 "curtain", "fridge", "freezer", "arabic series", "arabic movies",  
 "arab girls", "night life", "sony", "nokia", "mac", "dell", "acer", "compac",  
 "samsung", "toshiba", "sanyo", "japan", "china", "ticket", "reservation",  
 "thinking", "ERP", "oracle", "shakespeare", "britney spears", "shakira",  
 "angelina jolie", "brad pitt", "vandalism", "jackie chan",  
 "leonardo dicaprio", "vandalism", "comic", "famous", "email", "hotmail",  
 "healthy food", "angel", "stars", "hotel", "restaurant", "hostel", "inn",  
 "russia", "thesis", "english writing", "light show", "zara", "tommy", "CK",  
 "levis", "D&G", "diesel", "lee", "draft", "security", "real madrid",  
 "safety", "manchester", "argentina", "club", "pub", "bar", "pressure ulcer",  
 "risk", "risk assessment", "likert scale", "grading", "prevention", "recipe",  
 "intervention", "risk factors", "sore", "guidelines", "outlines", "general",  
 "viagra", "hottest women", "hottest songs", "hottest actress", "actor",  
 "actress", "hottest show", "latest tv shows", "eye", "nose", "mouth", "lips",  
 "teeth", "feet", "sand", "sandwich", "yogurt", "olive oil", "room", "lounge",  
 "flat", "most watched", "beginner", "senior", "junior", "consultant",  
 "fastest", "biggest", "smallest", "tallest", "easiest", "most eaten",  
 "most loved", "most beautiful", "most stupid", "stupidest", "most clever",  
 "bread", "english novel", "joyce", "ireland", "nationalism", "island",  
 "colony", "colonialism", "england", "arabian nights", "the orient",  
 "novelist", "zionism", "capitalism", "commonism", "liberalism", "critic",  
 "romance", "sadism", "danger", "most dangerous", "criteria", "tone", "tune",  
 "friend", "sister", "brother", "father", "mother", "grandfather", "family",

## APPENDIX A. THE EXPERIMENT SAMPLE DATA

---

```
"grandmother", "nephew", "niece", "view", "angle", "square", "triangle",
"rectangle", "circle", "watch", "clock", "timing", "hypothesis", "number",
"sample", "picnic", "journey", "trip", "allah", "mohammed", "prophet",
"advice", "basics of", "wikipedia", "distributed system", "conventional",
"convenient", "advantage", "disadvantage", "public", "private", "fact",
"mean", "name", "language", "morning", "body", "face", "map", "person",
"fine", "dark", "machine", "rest", "drive", "rain", "snow", "green", "road",
"street", "vitamin", "pharmacy", "medicine", "bottle", "bottle nick",
"battle", "africa", "asia", "europe", "union", "neighbor", "school", "agent",
"agency", "intelligence", "artificial", "artificial intelligence", "win",
"disables", "ability", "potential", "virus", "viral", "rival", "competition",
"detergent", "shopping", "teaching", "simultaneous", "crash", "tutoring",
"screen", "save", "pen", "stemming", "processing", "treatment", "cure",
"charge", "charger", "village", "country side", "background", "sensitive",
"weed", "cigarette", "incredible", "earthquack", "storm", "ocean", "whale",
"under water", "wild life", "aquarium", "hottest destinations", "up-to-date",
"innocent", "puzzle", "maze", "best technology", "2011 news", "best of 2011",
"top gear", "top horror movies of all time", "top action movies of all time",
"top romance movies of all time", "JSTL", "top thriller movies of all time",
"documentary", "chocolate", "oscar", "globe", "tattoo", "nipple", "piercing",
"juice", "upgrade", "install", "maintenance", "mechanic", "bluetooth", "LCD",
"infrared", "ultraviolet", "ring", "zipper", "heavy", "light", "thunder",
"palmtree", "best arab dates", "shelf", "carpenter", "pizza", "glue",
"saturday", "sunday", "monday", "tuesday", "wednesday", "thursday", "friday",
"january", "february", "march", "april", "may", "june", "july", "august",
"september", "october", "november", "december", "18 wheeler", "development",
"compose", "develope", "call", "survive", "story", "tail", "episode",
"flickr", "stubborn", "logic", "winning", "cheater", "cheating", "betray",
"illusion", "illuminate", "humiliate", "insulting", "bullying", "typing",
"way of thinking", "hottest research areas", "hottest technologies",
"bobel prize", "numaric system", "calculator", "alphabet", "best SMS", "end",
"how to start", "reading", "writing", "forgiveness", "forget", "try", "cut",
// ----- //

** The Italian Keywords **

// ----- //
"educazione", "tutorial", "ricerca", "studente", "accademia", "apprendimento",
"lezione", "Tecnologia", "sistema", "informatica", "programmazione", "scienza",
"musica", "pubblicita", "intrattenimento", "commedia", "foto", "stile",
"modello", "piano", "sociale", "storia", "mammifero", "legge", "beverage",
"bellezza", "arte", "disegno", "bello", "dipingere", "trasporto", "auto",
"treno", "volo", "viaggio", "turismo", "vacanza", "umano", "Persone",
"uomo", "ragazza", "creatura", "figli", "arabi", "cultura", "religione",
"ballare", "affari", "prodotto", "societa", "denaro", "economia", "ufficio",
"comunicazione", "linguaggio", "natura", "animale", "uccello", "pesce",
"giungla", "vita", "mondo", "salute", "ospedale", "militari", "alloggio",
"evento", "divertente", "triste", "cibo", "bere", "piatto", "ristorante",
" Sesso"
```

## A.2 Sample data statistics

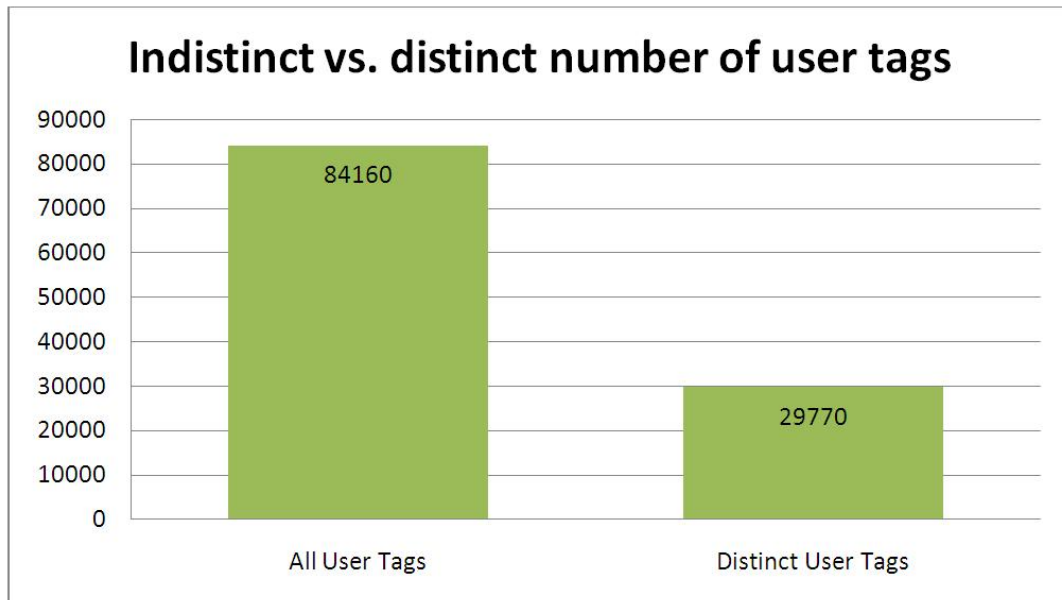


Figure A.1: The number of all user tags for the imported videos (indistinct) vs. the number of distinct user tags (the repetition of tags is omitted).

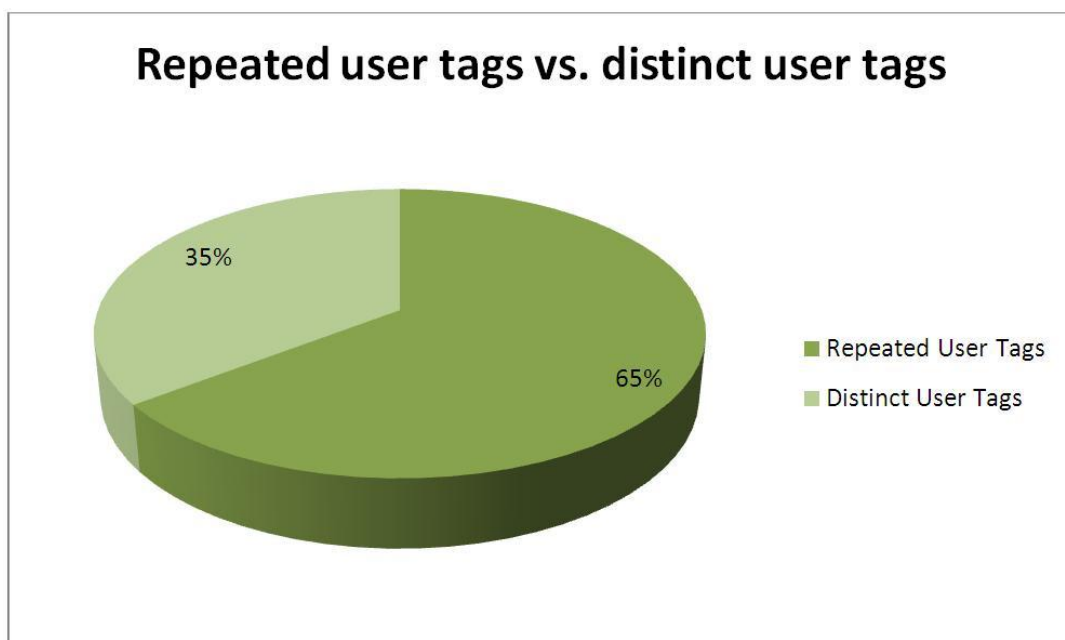


Figure A.2: The proportion of the distinct user tags to the repeated user tags. The total number of the tags represented in this chart is 84,160 (all the user tags).

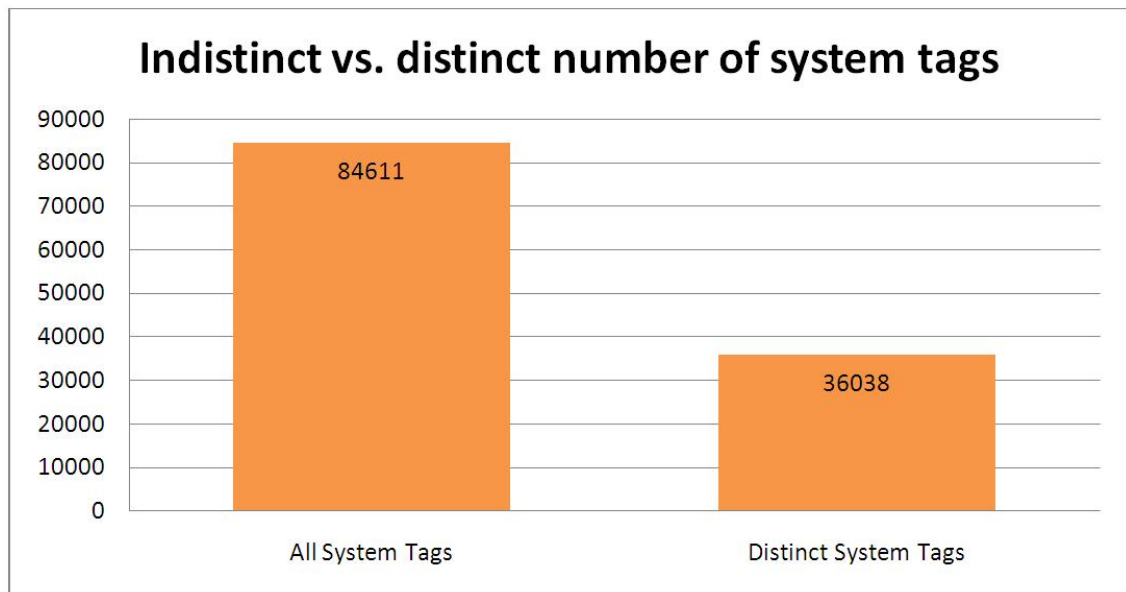


Figure A.3: The number of all system tags added to the imported videos (indistinct) vs. the number of distinct system tags (the repetition of tags is omitted).

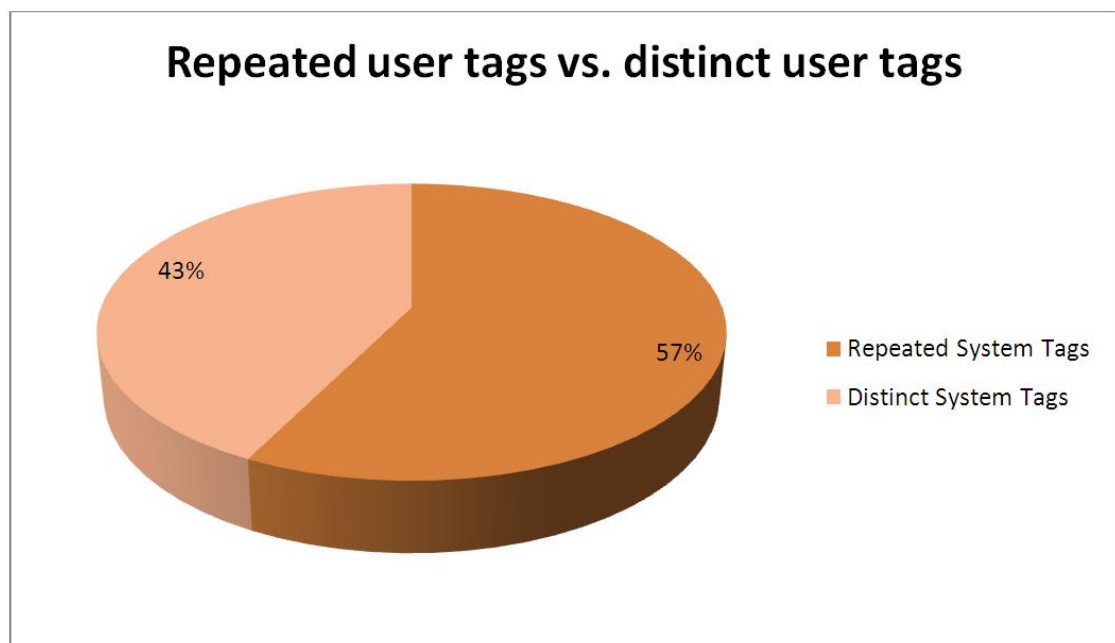


Figure A.4: The proportion of the distinct system tags to the repeated system tags. The total number of the tags represented in this chart is 84,611 (all the system tags).

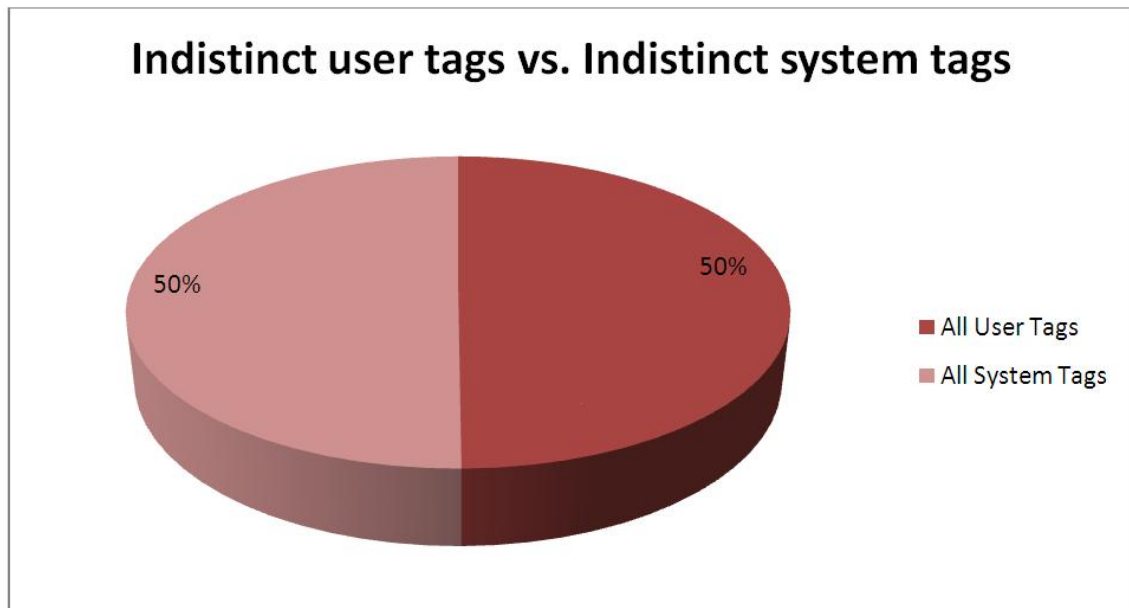


Figure A.5: The proportion of all the user tags to all the system tags stored in our database.

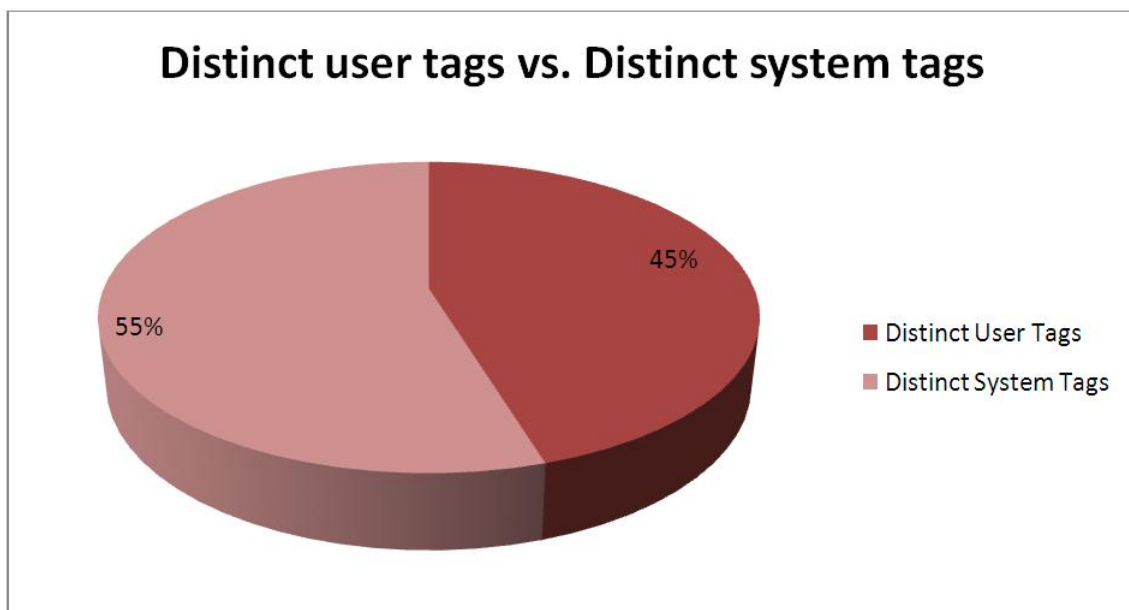


Figure A.6: The proportion of the distinct user tags to the distinct system tags stored in our database (the repetition of tags is omitted).

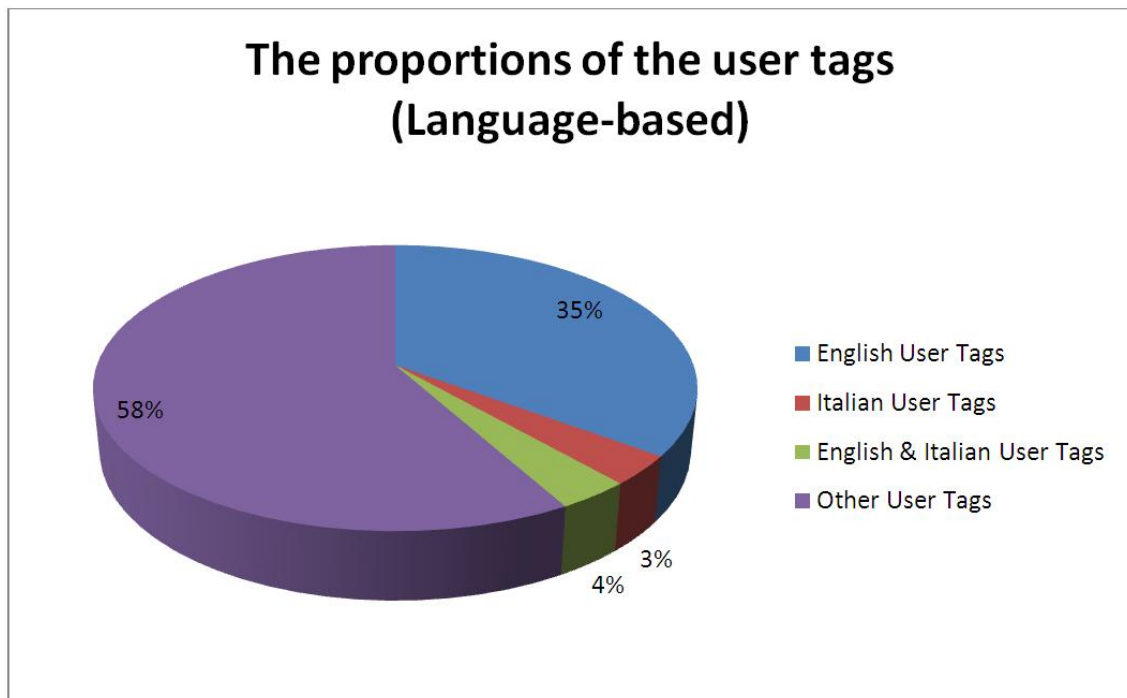


Figure A.7: The user tags are either English word, Italian words, words that are English and Italian at the same time, or shorthand writing tags.

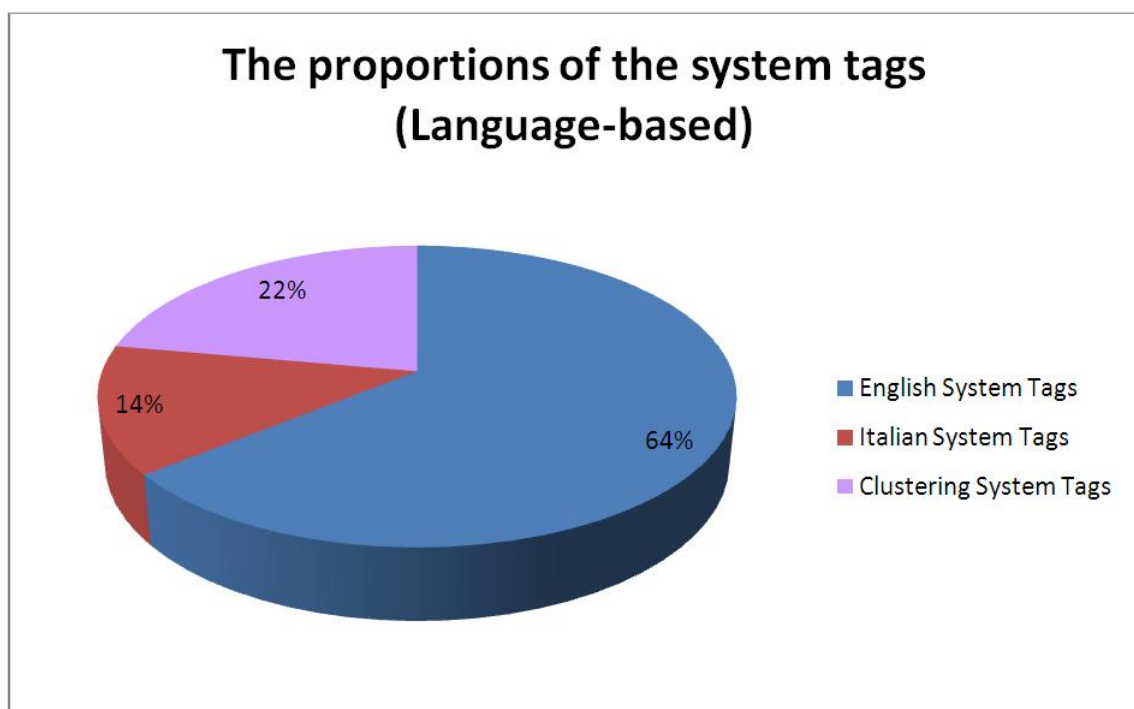


Figure A.8: The system tags are either English word, Italian words, or tags that came from clusters.

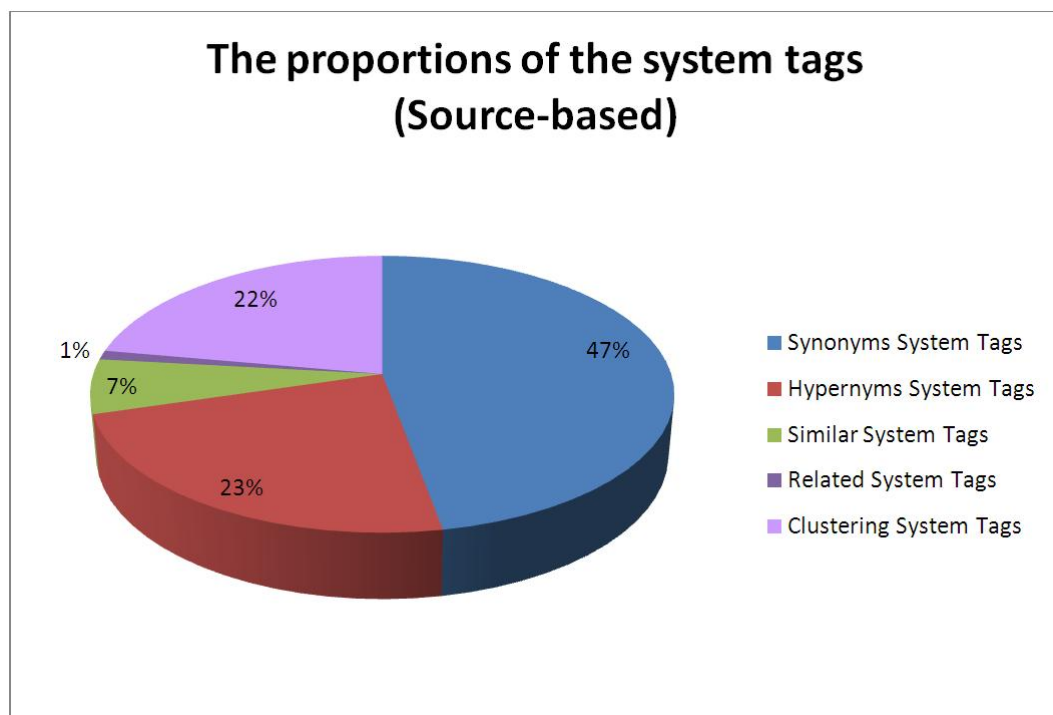


Figure A.9: The system tags came from either clusters or semantic relations. Both the English and Italian system tags came from synonymy relation, hypernymy relation, similar relation, or related relation.

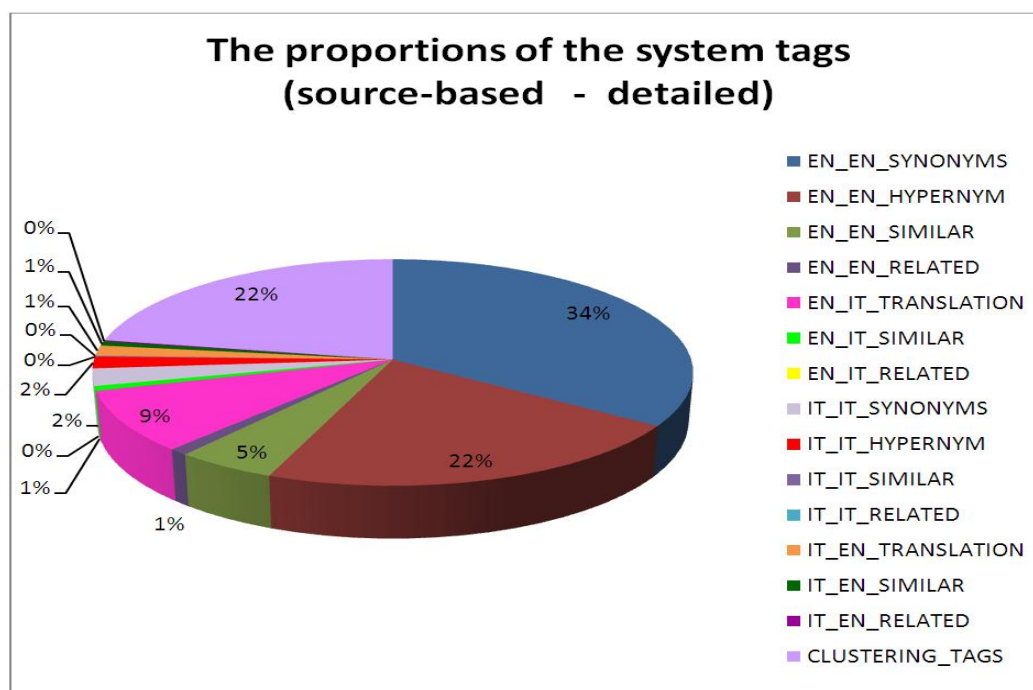


Figure A.10: Based on the language of the user tag and the language of the system tag, there are 14 sources of system tags (excluding the tags that came from clusters).



## Appendix B

### The Collected Data

#### B.1 The whole data set statistics

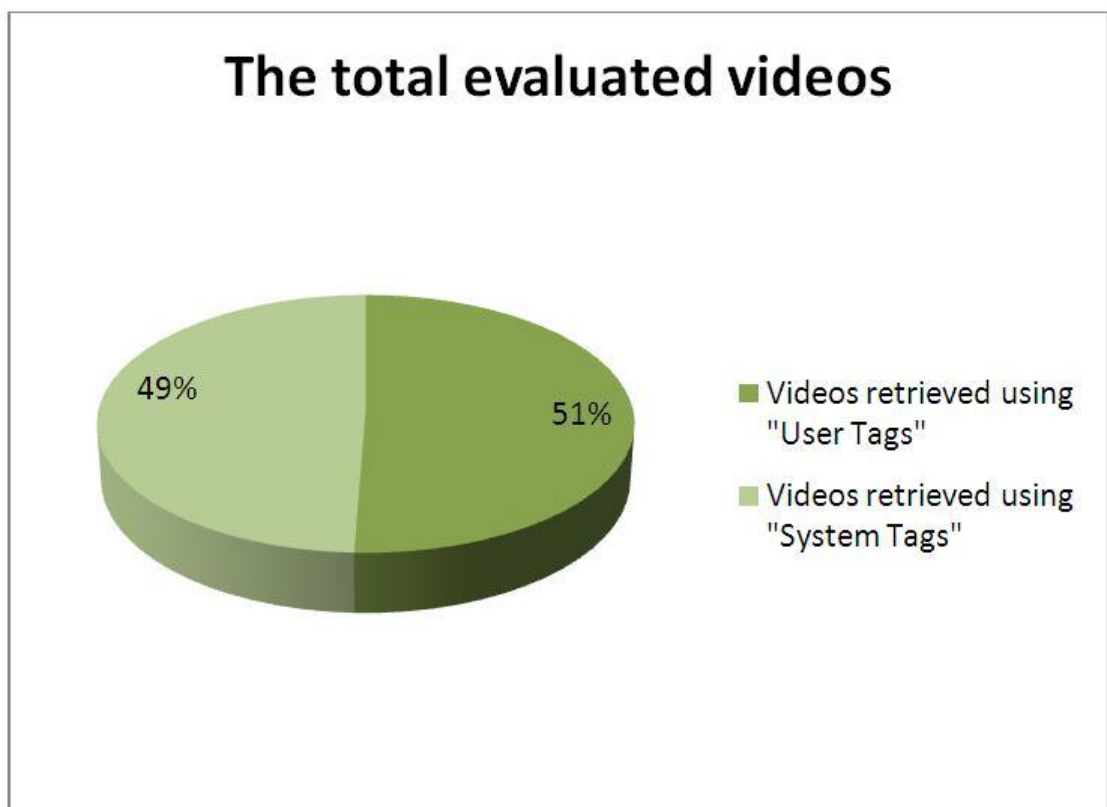


Figure B.1: The total evaluated videos (1391 videos).

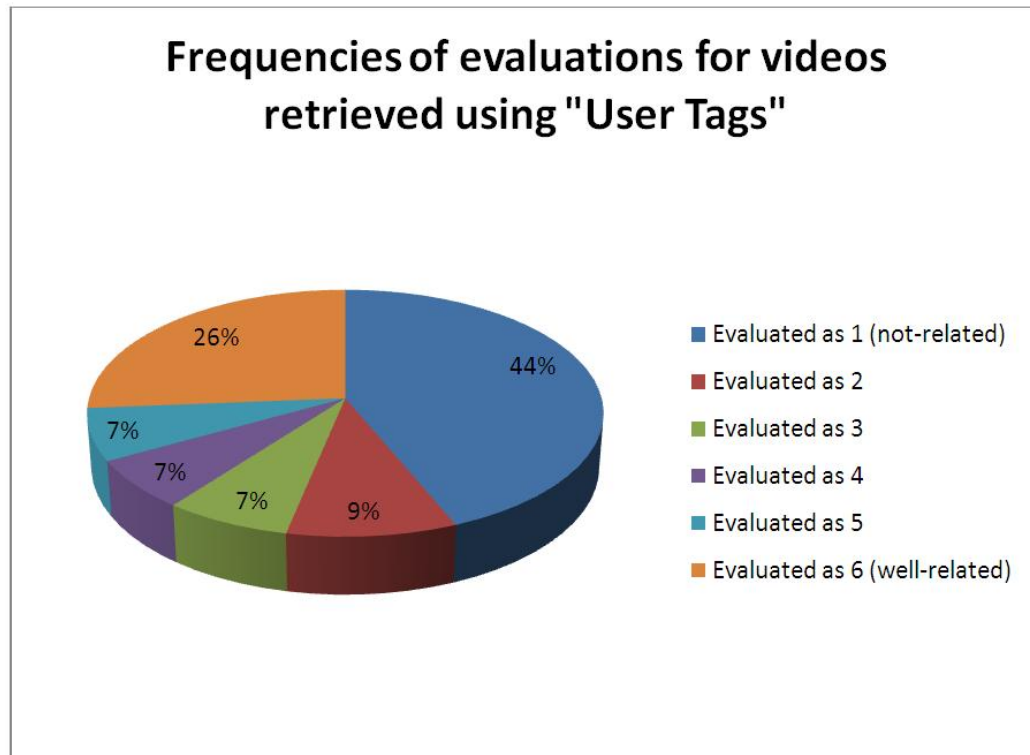


Figure B.2: Frequencies of evaluations for videos retrieved using “User Tags”.

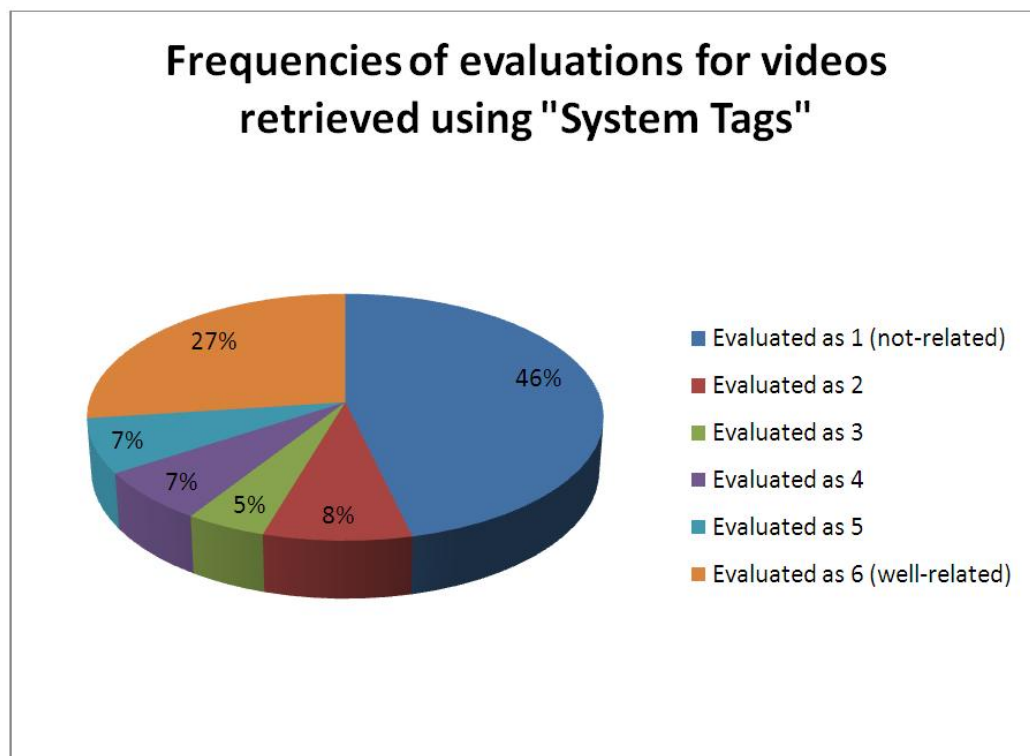


Figure B.3: Frequencies of evaluations for videos retrieved using “System Tags”.

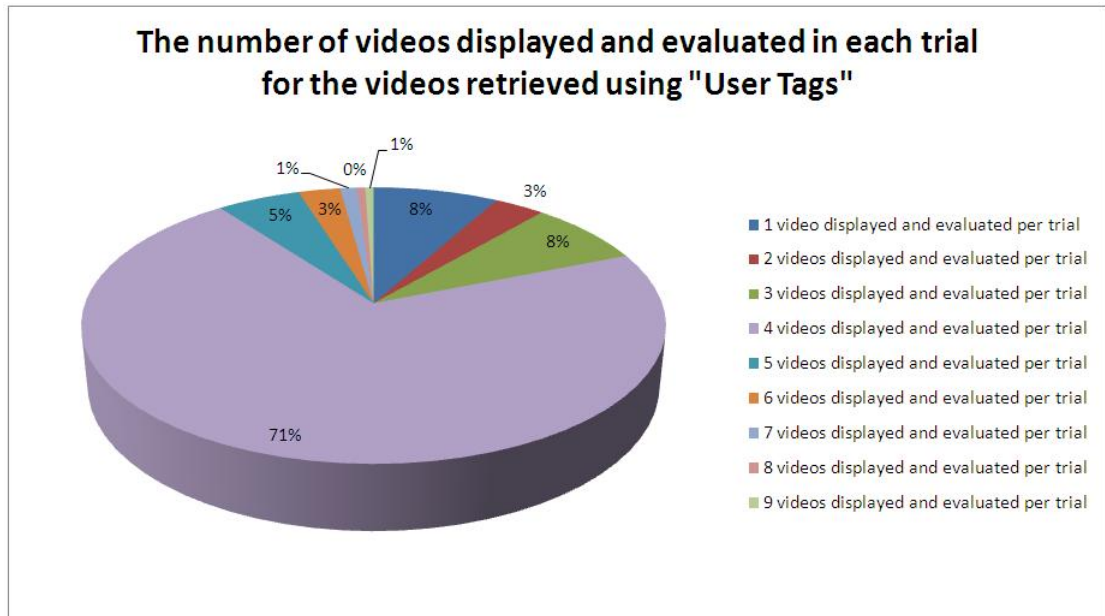


Figure B.4: The number of videos displayed and evaluated in each trial for the videos retrieved using “User Tags”.

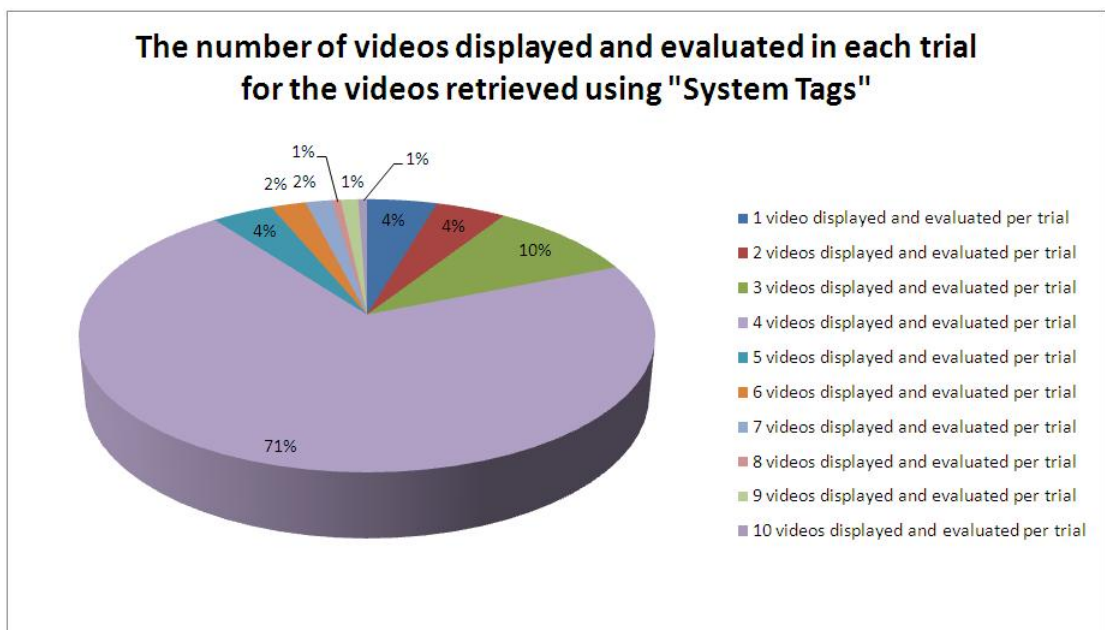


Figure B.5: The number of videos displayed and evaluated in each trial for the videos retrieved using “System Tags”.

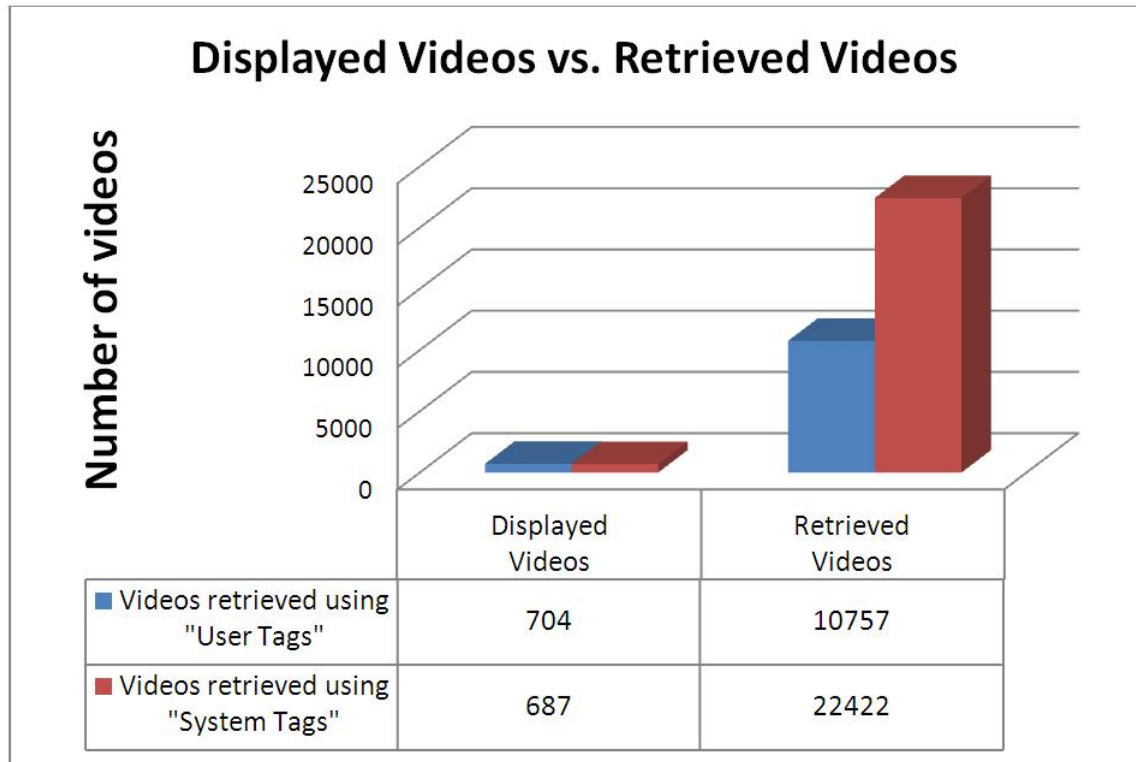


Figure B.6: Displayed Videos vs. Retrieved Videos.

## B.2 The collapsed data set statistics

	Frequency of collapsed evaluations for videos retrieved using <b>User Tags</b>	Frequency of collapsed evaluations for videos retrieved using <b>System Tags</b>
Collapsed to 1	425	406
Collapsed to 6	279	281
Sum	704	687

Table B.1: The frequencies of collapsed participants' evaluation for both groups.