

國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

標籤式使用者描述之語意關聯相似度研究

**A Comparative Study of Semantic Similarity of
Tag-based Profiles**

張琮傑

Tsung-Chieh Chang

指導教授：許永真 博士、陳文進 博士

Advisor: Jane Yung-jen Hsu, Ph.D.

Wen-Chin Chen, Ph.D.

中華民國九十八年六月

June, 2009



國立臺灣大學碩士學位論文
口試委員會審定書

標籤式使用者描述之語意關聯相似度研究

A Comparative Study of Semantic Similarity of Tag-based Profiles

本論文係張琮傑君（學號 R96922008）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 98 年 6 月 5 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

許永英 陳文進

（指導教授）

柯子弘

蔡宗翰

呂育道

系主任



Acknowledgments

要感謝許多人，有你們的幫助，我才能完成這篇碩士論文。

感謝我的指導教授，許永真老師，每次個別meeting時都讓我瞭解到哪裡還有不足該改進，還有什麼方向可以去嘗試看看。在口試倒數的最後關頭時，每天密集討論、修正，直到口試結束才告一段落；有了那段日子的改進，才使我的論文更為出色，進而在口試時能展現出更多成果。

感謝我的女朋友嘉涓，在我剛進入Semantic group時也是因為妳和怡靜才讓我決定研究tagging這個方向的問題。愈接近口試的日期，妳愈支持我、關心和鼓勵我，在準備口試投影片時也幫了我的忙並給我許多建議。謝謝妳嘉涓，在準備口試這段日子中充滿了緊張、焦慮，快要讓我喘不口氣來，幸好有妳陪著我，帶給我快樂和幸福。

感謝實驗室的夥伴們，David、小嫻、翰文、忠毅、好圓、小朋友、栗子、筱薇、元翔，很高興能和你們一起度過這兩年，這兩年中也有許多事感謝你們的幫忙才能順利完成，不管是準備口試時無數次的rehearsal，有許多人幫忙我找出了不少問題和建議改進的方法，或是David以及小嫻花了時間主辦各式活動和聚餐，或是和好圓、小朋友間的嘴炮科科之無聊對話，在研究生涯中增添不少笑聲。

感謝實驗室的學姊，不管是研究上或是需要任何物品或是跟跑公文流程等相關的事，只要問妳都能迅速得到詳細的答案，有妳的幫忙讓我們能專心於研究上，謝謝學姊。

感謝實驗室的其他人，不管是上一屆的學長姊，或是博班學長姊們，或是下一屆的學弟妹以及大學部專題生們，有你們讓我這兩年過得更愉快且順利，謝謝大家。

感謝我的好朋友們，雖然沒有時常聯絡，但是偶爾聚會吃吃喝喝打屁聊天就很棒，或是在BBS上聊個幾句問問近況也不錯，謝謝你們。

最後感謝我的父母，雖然不常把感謝等情緒表現出來，但在經濟上讓我無後顧之憂，使我不必另花心思去擔心學費等事。每次回台中時，也都能讓我徹底放鬆休息，暫時將其他事擱於一旁，謝謝。



Abstract

With the rapidly growing amount of information, especially in the era of Web 2.0, users experience the problem of information overload. Based on an accurate user profile, we can eliminate unwanted items and recommend the items to the user who interests. Though user profiles have been studied for a long time, constructing profiles based on *tags* is a new research topic which emerges in recent three years. Utilizing a user's set of tags to profile the user is reasonable because tagging associates an object with a set of words which represent the semantic concepts activated by the object from the user's perspective.

Nowadays, Common similarity measures between profiles just consider the same attributes only. But two tags may have semantic similarity even if they are not the same tag. In this thesis, we propose *semantic tag-based profiles* to enrich profiles based on *tag concepts* we proposed. Each tag concept is built from a core tag which connects other tags holding similar semantic meanings with the core tag. Furthermore, we propose an adaptive similarity measure for semantic tag-based profiles which integrates semantic similarity between tags.

Our evaluation is based on the data set crawled from *Delicious*, which is the most popular social bookmarking web site. The data set contains 20,578 users and 80,000 bookmarks after filtering the crawled data. From the results by empirical evaluation and user study, we show semantic tag-based profiles are better than tag-based profiles.



摘要

計算相似度(similarity)是研究上的熱門領域。以使用者為例，在計算使用者之間的相似度須先建立使用者描述(user profile)。在現今Web2.0的時代，使用者可以上傳自己的資料並用標籤(tag)管理；由於標籤是使用者對各個資料語意或概念上的描述，因此以標籤建立使用者描述可瞭解各使用者個人化的觀點與感興趣的主題。

目前計算使用者描述之間的相似度方法皆只考慮兩個使用者描述中共有的屬性。以標籤式使用者描述(tag-based user profile)為例，計算相似度時只考慮相同的標籤，字面上不同的標籤則會忽略不計。但是即使兩個標籤不同，以人類的知識會覺得它們之間具有語意相似度(semantic similarity)。因此在本論文中，我們將語意帶進標籤式使用者描述擴展成賦有語意的標籤式使用者描述(semantic tag-based user profile)，接著我們訂定衡量賦有語意的標籤式使用者描述之間的相似度方法。

我們的實驗資料來自於Delicious，它是目前資料量最豐富的社群書籤網站。我們共使用20,578位使用者以及80,000個網頁的資料來衡量我們提出的方法的效能。藉由研究上常用的評估方法以及我們設計的使用者調查，兩者皆顯示我們的方法較原本的標籤式使用者描述好。

Contents

Acknowledgments	iii
Abstract	v
List of Figures	xii
List of Tables	xiv
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Research Objectives	3
1.3 Thesis Structure	3
Chapter 2 Related Work	5
2.1 Social Tagging Systems	5
2.1.1 Tagging	5
2.1.2 Folksonomy	7

2.1.3	Usage Patterns	9
2.2	Semantic Similarity	10
2.2.1	Introduction	10
2.2.2	WordNet	13
2.2.3	ConceptNet	18
2.2.4	Web-based Approaches	22
2.3	User Profile	23
2.3.1	Demographic User Profile	24
2.3.2	Tag-based User Profile	24
2.4	Common Similarity Measures	27
2.4.1	Jaccard Coefficient	27
2.4.2	Cosine Similarity	27
2.4.3	Adjusted Cosine Similarity	28
2.4.4	Correlation-based Similarity	29
Chapter 3	Semantic Similarity Measure for Tag-based User Profiles	31
3.1	Background	32
3.1.1	Tag-based User Profile	32
3.1.2	Semantic Similarity	33
3.2	Semantic Similarity between Tag-based User Profiles	37
3.3	Proposed Solution	38
3.3.1	Semantic Tag-based User Profile	38

3.3.2	Similarity Measure for Semantic Tag-based User Profiles	39
Chapter 4	Methodology of Semantic Tag-Based User Profiles	41
4.1	Similarity Measure for Tag-based User Profiles	42
4.2	Semantic Tag-based User Profile	43
4.3	Similarity Measure for Semantic Tag-based User Profiles	47
4.3.1	Property of the Similarity Measure	50
Chapter 5	Experiment and Evaluation	53
5.1	Data Collection	53
5.2	Data Analysis	55
5.2.1	Data Filtering	55
5.2.2	Tag Coverages in Semantic Resources	57
5.2.3	Ratios of User's Tag Frequencies to Total Tag Frequency	60
5.3	Example Result	61
5.4	Empirical Evaluation	62
5.4.1	Precision-Recall Graph	63
5.4.2	Rank Accuracy Measures	65
5.5	User Study	68
5.5.1	User Study Design	69
5.5.2	User Study Result	70
Chapter 6	Conclusion	73

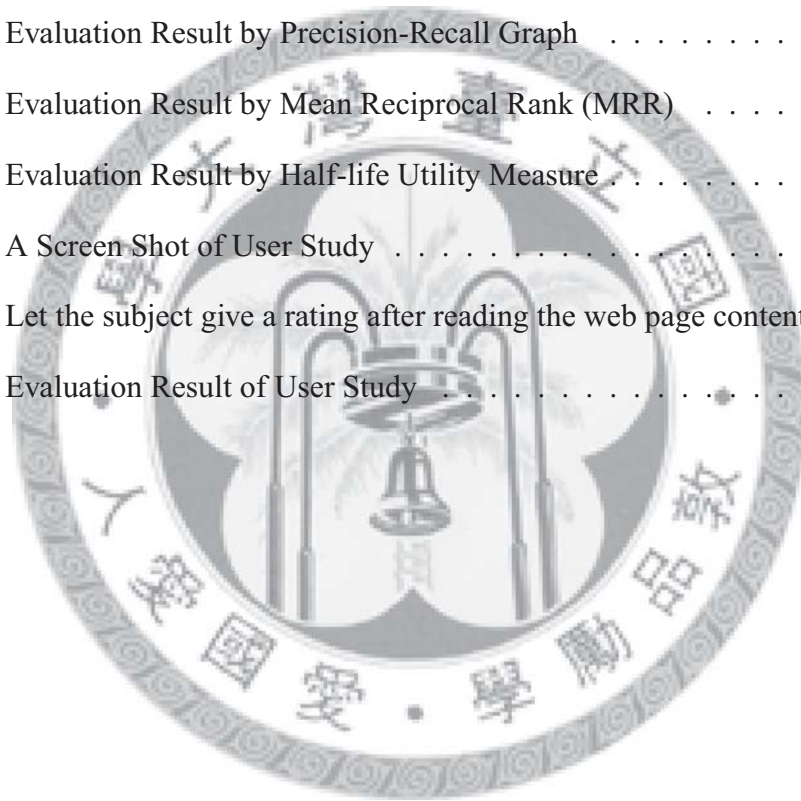
6.1 Summary of Contributions	74
6.2 Future Work	75
Bibliography	76



List of Figures

2.1	A ternary relation for representing an annotation which a user annotates a resource with a set of tags	8
2.2	The power-law distribution of the URL-saved frequency, the user-saved frequency, and the tag-used frequency [14]	10
2.3	A fragment of the semantic hierarchy of WordNet	14
2.4	ConceptNet represents assertions in the form of a semantic network.	19
4.1	Using cosine similarity to measure the similarity between tag-based profiles is not suitable	43
4.2	Construct a tag concept by spreading activation	46
4.3	There exists semantic relation between the tag concept of <code>driving</code> and the tag concept of <code>travel</code>	49
4.4	Select maximum similarity between among the pairs of one user's tag concept and the other user's all tag concepts	49

5.1	A user's data including his bookmark collection and used tags on Delicious	54
5.2	Distribution of the numbers of the users' total bookmarks	56
5.3	The Coverage of Tags in Semantic Resources	58
5.4	Average ratios of users' tag frequency of each rank to their total tag frequency	60
5.5	Evaluation Result by Precision-Recall Graph	64
5.6	Evaluation Result by Mean Reciprocal Rank (MRR)	66
5.7	Evaluation Result by Half-life Utility Measure	68
5.8	A Screen Shot of User Study	69
5.9	Let the subject give a rating after reading the web page content	70
5.10	Evaluation Result of User Study	71



List of Tables

2.1	A taxonomy of tagging motivations [2]	7
2.2	The benchmark data set for similarity measures	16
2.3	Evaluation results of semantic similarity methods	18
2.4	Partial of the types of the semantic relations in ConceptNet 3	21
5.1	The crawled data for evaluation	56
5.2	The list of top 20 tags ordered by frequency with their frequencies and the numbers of users used them	57
5.3	Example Result: Semantic Similarities between tag design and other tags	61

Chapter 1

Introduction

1.1 Motivation

The phenomenal rise of social media in recent years is transforming the average people from content readers to content publishers. People share a variety of media contents with their friends or the general public on social media sites. For example, people share bookmarks on Delicious¹, videos on Youtube², and photos on Flickr³. On social media sites, *tagging* is an important feature which enables people to easily add metadata to content, and these additional metadata can be used to improve search mechanisms or better structure the data for browsing.

On social networking sites, users who we are *familiar* with are often the ones who we share valuable information with; on social media recommender sites, users with

¹<http://delicious.com>

²<http://www.youtube.com>

³<http://www.flickr.com>

similar tastes are often the ones who provide recommendations to help us make better choices. Regardless of the type of connection between people involved, at the heart of developing these systems is an attempt to identify *overlap* between user profiles that appropriately reflect the preference and behavior of the user.

However, a typical personal profile consisting of simple demographic data, such as the name, affiliation, or interests, provides an inadequate description of the individual, as they are often *incomplete*, mostly *subjective* and cannot reflect dynamic changes. Tagging is fundamentally about sense-making which is a process in which information is categorized and labeled and, critically, through which meaning emerges. Observing that the rich online media collected by an individual provide important insights about the person, we capitalize on such data by profiling a user with an aggregation of tags associated with his social media.

Accurate profiling of a user allows system developers to provide personalized services such as more precise information filtering and more accurate information retrieval results. Yet identifying overlapping connections between users based on their profiles allows for the design of a wider range of more advanced services to be offered. In real life, connections with the right people often allow us to have a competitive advantage over others, whether it be getting a job offer, developing a sales strategy, or simply seeking for a good advice. Similar scenarios are observed in the online world. For example, collaborative filtering recommender systems [1] draw on the similarity between user ratings to make predictive recommendations. With connecting to the more similar users, the recommended items could attract the target user more.

1.2 Research Objectives

In this thesis we propose the semantic tag-based user profile to represent a user's interests particularly. We believe that the profile enriched by semantic relations between tags better reflect the preferences and knowledge of the user. In addition, we propose a similarity measure for semantic tag-based profiles to solve the problem on calculating the similarity between tag-based profiles by cosine similarity. In the absence or sparsity of rating information, similarity between semantic tag-based profiles can provide more diversified and more serendipitous recommendation results. Furthermore, similarity relations between users allows for the construction of a social network structure, on which techniques in social network analysis may be applied to observe and fine-tune the overall evolving system.

1.3 Thesis Structure

In what follows, we will start by briefly reviewing related research in social tagging systems, semantic similarity including WordNet [22] and ConceptNet [8] with related similarity measures [15, 30], and user profiling with similarity measures between profiles. In Chapter 3, we then formally define a tag-based user profile, a tag concept, and a semantic tag-based user profile consisting of a set of tag concepts with a similarity measure for semantic tag-based profiles sequentially. In Chapter 4, we first give a synopsis of how a tag-based profile can be constructed. Then we propose approaches to measure semantic similarities between tags based on WordNet, ConceptNet, or Google snippets. These approaches are fundamental to the construction of a tag concept, and

a semantic tag-based user profile consists of a set of tag concepts. Then we propose an approach to measure a similarity between semantic tag-based user profiles. In Chapter 5, We introduce the data set we crawled and the analysis of the data. We construct three semantic tag-based user profiles for a user, which are based on WordNet, ConceptNet, and on Google snippets separately. The baseline is a tag-based user profile for the same user to compare with our proposed approaches by 5-fold cross evaluation. We also have a user study for evaluation. Finally, we express a summary of this thesis in Chapter 6.



Chapter 2

Related Work

In this chapter, we present a brief introduction of recent researches related with tagging, semantic similarity with related resources, and user profiling.

2.1 Social Tagging Systems

2.1.1 Tagging

Tagging is commonly used on social media sites to add comments about the media content, or to help organize and retrieve relevant items. Tagging associates an object with a set of words, which represent the semantic concepts activated by the object at the cognitive level. While categorization is a primarily subjective decision process, tagging is a social indexing process.

Web 2.0 web sites allow users to do more than just retrieve information, and tagging

is one of the supplied services for users. Delicious¹ (del.icio.us formerly) is the most popular social bookmarking web service site for storing, sharing, and discovering web bookmarks. The first version was published in 2003 and now it has more than 5 million users and 180 million unique URLs². Each user can save, manage and share web pages online without restricting to one personal computer only. A user can tag each URL he liked with freely chosen terms and then save it in his bookmark collection. Later, the user can retrieve all bookmarks tagged by a specific term. Furthermore, he also can acquire the bookmarks tagged by other users with the same tag. Users also can see the “hotlist” of bookmarks from the homepage of Delicious. “Popular” and “recent” pages are also existed for users to discover the useful and interesting bookmarks they like.

Users have to spend additional time thinking and annotating their items with suitable tags, so why do users tag? Ames and Naaman [2] made a user study of ZoneTag³ users, which is a camera-phone application used to upload photos taken by the phone to Flickr⁴, which is the biggest image hosting web site. They offered a simple taxonomy of motivations for tagging along two dimensions, *sociality* (which includes *self, social*) and *function* (which includes *organization, communication*), described in Table 2.1. By interviewing the participants, they suggested that most of our participants were motivated to tag by organization for the general public (search, self-promotion), with self-organization (for later retrieval) and social communication (for friends, family, and the public) tied for second.

¹<http://delicious.com>

²<http://blog.delicious.com/blog/2008/11/delicious-is-5.html>

³<http://zonetag.research.yahoo.com>

⁴<http://www.flickr.com>

		Function	
		Organization	Communication
Sociality	Self	Retrieval Search	Context for self Memory
	Social	Contribution, attention Ad hoc photo pooling	Content descriptors Social Signaling

Table 2.1: A taxonomy of tagging motivations [2]

2.1.2 Folksonomy

Differing from formal taxonomies and classification schemes, social tagging systems lack a predefined terming structure. They rely on shared and emergent social structures and behaviors, as well as related conceptual and linguistic structures of the user community. The term folksonomy is usually used to refer to the structure of tags in these tagging systems.

Collaborative tagging systems allow users to choose their own words as tags to describe their favourite Web resources, resulting in an emerging classification scheme now commonly known as a folksonomy.

Folksonomies are user-contributed data aggregated by collaborative tagging systems. In these systems, users are allowed to choose terms freely to describe their favourite Web resources. A folksonomy is generally considered to consist of at least three sets of elements, namely users, tags and resources.

To model networks of folksonomies at an abstract level, Peter Mika [21] represented such a system as a tripartite graph with hyperedges by extending the traditional bipartite model of ontologies (concepts and instances) by incorporating actors in the model. The set of vertices is partitioned into the three (possibly empty) disjoint sets

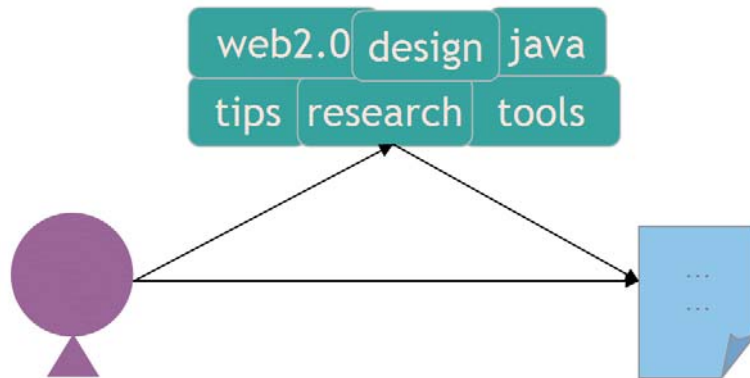


Figure 2.1: A ternary relation for representing an annotation which a user annotates a resource with a set of tags

$A = \{a_1, \dots, a_k\}$, $C = \{c_1, \dots, c_l\}$, $I = \{i_1, \dots, i_m\}$ corresponding the set of actors (users), the set of concepts (tags, keywords) and the set of annotated resources (bookmarks, photos etc.).

Leveraging the data of a folksonomy, Jäschke *et al.* [11] introduced the FolkRank algorithm, which computes a topic-specific ranking of the elements in a folksonomy, and defeated collaborative filtering algorithms in the area of recommender systems [1]. FolkRank needs a preference vector to determine the topic, and it may have any distribution of weights. Typically a single entry or a small set of entries is set to a high value, and the remaining weight is equally distributed over the other entries. And a topic can be defined in the preference vector not only by assigning higher weights to specific tags, but also to specific resources and users. These three dimensions can even be combined in a mixed vector. Similarly, the ranking is not restricted to resources, it may as well be applied to tags and to users.

2.1.3 Usage Patterns

Tagging has been studied by researchers in recent years, and some common patterns of collaborative tagging are revealed. Golder and Huberman [6] found that many bookmarks in Delicious reach their peak of popularity as soon as they reach Delicious, and some bookmarks are “rediscovered” and then experience a rapid jump in popularity after a long time. They also found that the frequency of a tag is a nearly fixed proportion of the total frequency of all tags used in a bookmark after the first 100 users empirically. This stability has important implications for the collective usefulness of individual tagging behavior.

Power-law distribution is also an important observation on tagging systems. A power law is a relationship between two scalar quantities x and y of the form:

$$y = cx^\alpha \quad (2.1)$$

where α and c are constants characterizing the given power law. Without loss of generality, Eq. 2.1 can also be written as:

$$\log y = \alpha \log x + \log c \quad (2.2)$$

In the form of Eq. 2.2, a fundamental property of power-law becomes apparent. which power laws are straight lines when plotted in log-log space. Halpin *et al.* [7] found that relative position of a tag ordered by used frequency in a bookmark and number of times the tag is used are power-law relationship. In the data set with 1.4 million URLs collected from Li *et al.* [14], they also observed power-law distributions of the URL-saved frequency, the user-saved frequency, and the tag-used frequency.

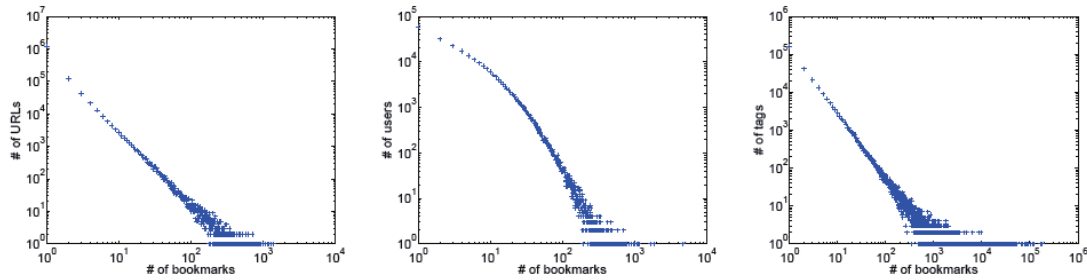


Figure 2.2: The power-law distribution of the URL-saved frequency, the user-saved frequency, and the tag-used frequency [14]

2.2 Semantic Similarity

2.2.1 Introduction

From psychological experiments, Douglas L. Medin *et al.* [19] showed that semantic similarity is context-dependent. For example, a snake and a parrot were judged much less similar when no explicit context was given than when the context of *pets* was provided. For another example, if the context is “the outside covering of living objects,” then *skin* and *bark* are more similar than *skin* and *hair*; however, the opposite is true if the given context is body parts.

They also proposed that semantic similarity may be asymmetric with respect to direction of similarity comparison. To say that *surgeons* are like *butchers* means something different than to say *butchers* are like *surgeons*. The former criticizes surgeons and the latter compliments butchers. Nevertheless, experimental results about investigating the effects of asymmetry suggested that the average difference in ratings for a word pair is less than 5 percent.

Two different strategies have been tried calculating semantic similarity. One is based on **co-occurrence**, and the other is on **substitutability**. Syntagmatic word associations, which arise from the co-occurrence of words in discourse and are attributed to association by contiguity. And paradigmatic word associations, which arise from the substitutability of words in discourse and are attributed to mediated association, i.e. to associations mediated by common contexts.

Consider the first strategy based on co-occurrence:

1. List all the words that occur in a set of contexts of item A.
2. List all the words that occur in a comparable set of contexts of item B.
3. Calculate some normalised coefficient representing the proportion of words common to the two lists.

The more likely it is that words co-occurring with A also co-occur with B, the more similar the two sets of contexts are judged to be.

An advantage of measures based on co-occurrence is that they are easily calculated with the help of modern computers, but from Rubenstein and Goodenough's viewpoint [26], this measure of contextual similarity confirmed the contextual hypothesis only for short distances in semantic space.

The second strategy is based on substitutability:

1. Collect a set of sentences using item A.
2. Collect a set of sentences using item B.

3. Delete A and B, shuffle the resulting contexts.
4. Challenge subjects to sort out which is which.

The more contexts there are that will take either item, the more similar the two sets of contexts are judged to be.

This approach for estimating the similarity of sets of contexts has been called “the method of sorting”. A subject’s task is to arrange sets of linguistic contexts for two (or more) words into groups of contexts all capable of accepting the same missing word. If two words were perfect synonyms, it would be impossible to discriminate the contexts of one from the contexts of the other.

The problem with co-occurrence measures is not merely that they dismember the contexts they are supposed to represent. A more serious problem is that they do not approach these tasks the way people do - whatever a word’s contextual representation may be, it is certainly not a collection of other words. If the argument advanced here is correct, people’s knowledge of how to use a word is organized to enable them to recognize rapidly the contexts it goes into. Consequently, measures of contextual similarity based on substitutability come closer to the desired goal. But the disadvantage of measures based on substitutability is that there is no quick and easy computer algorithm for calculating them.

Tags are composed of words that have inherent semantic meanings. In the next subsections we introduce two different semantic resources, WordNet⁵ and ConceptNet⁶, and approaches proposed by researchers for measuring semantic similarity between

⁵<http://wordnet.princeton.edu>

⁶<http://conceptnet.media.mit.edu>

words.

2.2.2 WordNet

Introduction

WordNet [22] is arguably the most popular and widely used semantic resource in the computational linguistics community today. It groups English words into sets of synonyms called synsets, provides short and general definitions, and records the various semantic relations between the synsets. As of 2006, the database of the newest version 3.0 contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs.

WordNet partitions the lexicon into nouns, verbs, adjectives, and adverbs. Nouns, verbs, adjectives, and adverbs are organized into synonym sets, called *synsets*. A synset represents a concept in which all words have similar meaning. Thus, words in a synset are interchangeable in some syntax. Knowledge in a synset includes the definition of these words as well as pointers to other related synsets.

WordNet is organized by semantic relations. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. Most relations in WordNet are “*is a*” (IS-A) relations, and relations are constructed in a hierarchic structure as in Fig. 2.3. The IS-A hierarchical structure of the knowledge base is important in determining the semantic distance between words, and researchers apply the attributes from the hierarchical structures in the functions for calculating semantic similarity. Otherwise,

they also retrieve *information contents* of each word from corpora and regard them as parameters in similarity functions.

For evaluating performances of similarity functions researchers invented, they needed a benchmark for comparing with other researchers' results fair. In the following we will first introduce the benchmark data set briefly and related approaches for calculating semantic similarity.

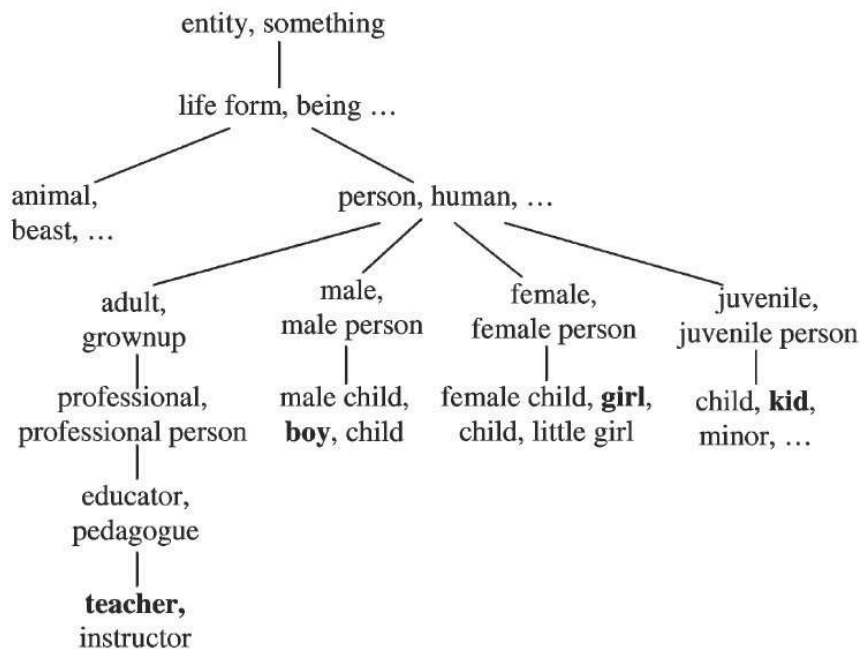


Figure 2.3: A fragment of the semantic hierarchy of WordNet

The Benchmark Data Set for Similarity Measures

George A. Miller said, “What people know when they say that they know a word is not how to recite its dictionary definition - they know how to use it (when to produce

it and how to understand it) in everyday discourse.” We can know explanation about various meanings of a word, but we cannot know similarity between words from the definition.

Semantic similarity is usually estimated by asking people to rate pairs of words with respect to their likeness of meaning. The first benchmark data set was built by Herbert Rubenstein and John B. Goodenough in 1965 [26]. The data set contained 65 pairs of ordinary English nouns originally for synonymy judgment. 51 undergraduate subjects were asked for judging similarity between each pairs with a value from 0.0 to 4.0, where 0.0 represents no similarity of meaning and 4.0 perfect synonymy.

In 1991, George A. Miller *et al.* [23] reproduced the experiment described above. From the result of the experiment, three sections were sectioned by similarity value of word pairs, including the high level between 3 and 4, the intermediate level between 1 and 3, and the low level between 0 and 1. They selected 10 pairs of nouns from each section, 30 pairs of nouns from the original list totally. 38 undergraduates were paid to serve as subjects for rating each pair of nouns with the same range of similarity value.

The result are listed in Table 2.2 with those corresponding similarity values obtained by Rubenstein and Goodenough in 1965. The two sets of ratings were in good correspondence, and the Pearson correlation coefficient is 0.97. It means that people are not only able to agree reasonably well about the semantic distances between concepts, but their average estimates remain remarkably stable over 26 years.

Noun Pair	Miller [23]	Rubenstein [26]
car - automobile	3.92	3.92
gem - jewel	3.84	3.94
journey - voyage	3.84	3.58
boy - lad	3.76	3.82
coast - shore	3.70	3.60
asylum - madhouse	3.61	3.04
magician - wizard	3.50	3.21
midday - noon	3.42	3.94
furnace - stove	3.11	3.11
food - fruit	3.08	2.69
bird - cock	3.05	2.63
bird - crane	2.97	2.63
tool - implement	2.95	3.66
brother - monk	2.82	2.74
lad - brother	1.66	2.41
crane - implement	1.68	2.37
journey - car	1.16	1.55
monk - oracle	1.10	0.91
cemetery - woodland	0.95	1.18
food - rooster	0.89	1.09
coast - hill	0.87	1.26
forest - graveyard	0.84	1.00
shore - woodland	0.63	0.90
monk - slave	0.55	0.57
coast - forest	0.42	0.85
lad - wizard	0.42	0.99
chord - smile	0.13	0.02
glass - magician	0.11	0.44
rooster - voyage	0.08	0.04
noon - string	0.08	0.04

Note: Mean ratings on a scale from 0 to 4 by 38 subjects in Experiment I (Oswego) compared with mean ratings reported by Rubenstein and Goodenough (R and G) of 30 noun pairs.

Table 2.2: The benchmark data set for similarity measures

Categories of Approaches for Measuring Semantic Similarity

Several methods for determining semantic similarity between words have been proposed in the literature and most of them have been tested and on WordNet. Semantic similarity methods are classified into four main categories:

- Edge Counting Methods: Measure the similarity between two words (concepts) as a function of the length of the path linking the words and on the position of the words and their subsumer in the taxonomy.
- Information Content Methods: Measure the difference in information content of the two words as a function of their probability of occurrence in a corpus. This approach is independent of the corpus and also guarantees that the information content of each word is less than the information content of its subsumed words. [25]
- Feature-based Methods: Measure the similarity between two words as a function of their properties (e.g., their definitions or “losses” in WordNet) or based on their relationships to other similar words in the taxonomy.
- Hybrid methods: Combine the above ideas.

G. Varelas *et al.* [32] presented a comparative evaluation of various semantic similarity methods. In accordance with previous research, they evaluated the results obtained by applying the semantic similarity methods invented by former researchers, and then calculated the Pearson correlation coefficient between the result of each method and the rating values from Miller’s experiment [23]. The results are listed in Table 2.3.

Method	Category	Correlation
Wu [33]	Edge Counting	0.74
Li [15]	Edge Counting	0.89
Resnik [25]	Information Content	0.79
Tversky [31]	Feature	0.73
Jiang [12]	Hybrid	0.83

Table 2.3: Evaluation results of semantic similarity methods

From the results in Table 2.3, the feature-based method from Tversky [31] gives the poorest performance against human ratings. Resnik’s [25] information content method provides a better similarity measure with a correlation of 0.79. Jiang and Conrath [12] proposed a hybrid method, and they combined information content with edge counting using a formula that also took into consideration local density, node depth, and link type, which obtained a correlation of 0.83. The best result is Li *et al.*’s method [15], and we will describe the method in Sec. 3.1.2.

2.2.3 ConceptNet

Introduction

Commonsense knowledge collects human experience and encompasses knowledge about different aspects of typical everyday life.

The Open Mind Common Sense (OMCS) project created by MIT Media Lab in 2000 serves as a distributed solution to the problem of common sense acquisition, by enabling the general public to enter common sense into the system with no special training or knowledge of computer science. The project currently has 14,000 registered English language contributors.

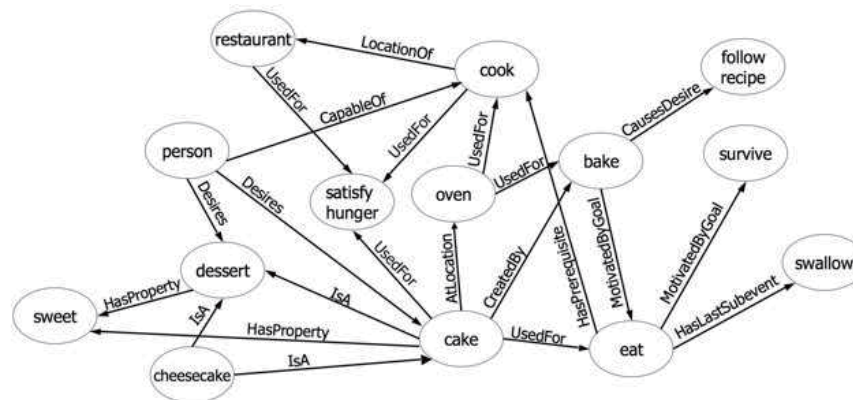


Figure 2.4: ConceptNet represents assertions in the form of a semantic network.

OMCS collects data over 700,000 assertions of commonsense knowledge by interacting with its contributors in activities which elicit different types of common sense knowledge. Some of the data is entered free-form, and some was collected using semi-structured frames where contributors were given assertions and would fill in a word or phrase that completed the assertion. For example, given the frame “___ can be used to ___.”, one could fill in “a pen” and “write”, or more complex phrases such as “take the dog for a walk” and “get exercise”.

ConceptNet [16, 8] is a representation of the Open Mind Common Sense corpus described above. From the semi-structured English assertions in OMCS, they extract knowledge and mine it into a semantic network. It has 21 relation-types that describe different relations among things, events, characters, etc.

Fundamental Elements

Whereas WordNet excels at lexical reasoning, the benefit of ConceptNet is contextual commonsense reasoning. ConceptNet is designed to be use as a natural-language-processing tool-kit which supports many practical textual-reasoning tasks including topic-gisting, analogy-making, and other context oriented inferences. In the newest version, ConceptNet 3 [8], developers focus on the usefulness of the data in the OMCS project and modularize ConceptNet for using other data sets easily.

The basic nodes of ConceptNet are *concepts*, which are aspects of the world that people would talk about in natural language. Concepts correspond to selected elements of the common-sense assertions that users have entered, and they can represent noun phrases, verb phrases, and adjective phrases. Concepts tend to represent verbs only in complete verb phrases, so “go to the store” and “go home” are more typical concepts than the simple verb “go”.

In a semantic network where concepts are the nodes, the edges are *predicates*, which express relationships between two concepts. Predicates are extracted from the natural language assertions that contributors entered, and express types of semantic relationships such as *ISA*, *PartOf*, *LocationOf*, and *UsedFor*. Now there are 21 basic relation types, and Havasi *et al.* [8] planned to add more in the future.

After comparing each assertion with an ordered list of patterns which represent sentence structures that are commonly used to express the various relation types in ConceptNet, the result is “raw predicates” that relates two strings of text. Table 2.4 shows some examples of patterns that express different relations. The phrases that fill

Relation	Example sentence pattern
IsA	NP is a kind of NP.
MadeOf	NP is made of NP.
UsedFor	NP is used for VP.
CapableOf	NP can VP.
DesireOf	NP wants to VP.
CreatedBy	You make NP by VP.
InstanceOf	An example of NP is NP.
PartOf	NP is part of NP.
PropertyOf	NP is AP.
EffectOf	The effect of VP is NP VP.

Table 2.4: Partial of the types of the semantic relations in ConceptNet 3

the slots in a pattern are the phrases that will be turned into concepts. The normalization process, including removing punctuation, stop words and stemming, determines which two concepts these strings correspond to, turning the raw predicate into a true edge of ConceptNet.

The assertions that currently comprise ConceptNet were collected from the Open Mind Common Sense web site, which used prompts such as “What is one reason that you would *ride a bicycle*?” to collect assertions of common sense from its users. If concept X and concept Y appear in corresponding places in many equivalent predicates, they are considered to be similar concepts. Then, if concept X appears in a predicate that is not known about concept Y, Open Mind Commons can hypothesize that the same predicate is true for Y, and it can make this inference stronger by finding other similar concepts that lead to the same hypothesis.

2.2.4 Web-based Approaches

Despite the usefulness of semantic resources created by experts or folks, plentiful information on the Web is also an ideal resources for measuring semantic similarity between words or texts. With utilizing powerful web search engine, we can retrieve the web pages with reliable qualities from billions of web pages in the world.

Web-based approaches are for measuring the similarity between such short text snippets that captures more of the semantic context of the snippets rather than simply measuring their term-wise similarity. To achieve this goal, we can leverage the large volume of documents on the web to determine greater context for a short text snippet. By examining documents that contain the text snippet terms we can discover other contextual terms that help to provide a greater context for the original snippet and potentially resolve ambiguity in the use of terms with multiple meanings.

This kind of approaches is based on *query expansion* techniques [28], which have long been used in the Information Retrieval community. Such methods automatically augment a user query with additional terms based on documents that are retrieved in response to the initial user query or by using an available thesaurus. However, the usage of query expansion for measuring semantic similarity between words or texts differs from the previous work.

Referring to the approach proposed by Sahami and Heilman [27], the traditional goal of query expansion has been to improve recall (potentially at the expense of precision) in a retrieval task. But their focus is on using such expansions to provide a richer representation for a short text in order to potentially compare it robustly with

other short texts. Moreover, traditional expansion is focused on creating a new query for retrieval rather than doing pair-wise comparisons between short texts. Thus, the web-based approaches are quite different than the use of query expansion in a standard Information Retrieval context.

After retrieving enough search snippets returned from web search engine, Sahami and Heilman [27] converted each snippets into a document vector, and then combined and normalized those vectors into one vector. For measuring semantic similarity between two texts, we can measure the cosine similarity of the two corresponding vectors.

2.3 User Profile

With the rapidly growing amount of information, especially on the web, users are often overwhelmed by the large amount of information they have to go through and experience the problem of information overload. Information overload is a situation whereby the individual is no longer able to effectively process the amount of information he or she is exposed to.

Generation of user profiles from samples of user interests and characteristics is a hot topic for research because user profiles can be used to retrieve resources matching user interests. A common application takes sample data (documents) that a user finds interesting (or uninteresting) and generates a user profile of the user's interests. If a user profile is generated exactly, we can filter and ignore unwanted items, or find out and recommend items to the user he probably likes.

2.3.1 Demographic User Profile

Krulwich [13] developed an approach to the task of user profiling called demographic generalization. He classified users in terms of users' demographic data from a commercially available database that encompasses the interests of people, and these classifications are used as general characterizations of the users and their interests. If only one cluster matches, all the data available for the cluster are used as a broad profile of the user, and the process ends. If more than one cluster matches the user data, the demographic variables whose values are similar in all the matching clusters form a partial profile of the user.

2.3.2 Tag-based User Profile

From a research perspective, the literature on collaborative tagging is rapidly expanding, and tag-based user profile is a new research topic in recent two years.

E. Michlmayr *et al.* [20] created user profiles from tagging data of users' bookmark collections. Each bookmark in the collection is composed of a title, a description, a URL, a bookmarked date, and a set of tags usually. For creating the profile, they focused on the tags and their temporal ordering by increasing date. Three approaches are proposed by them, including naive approach, co-occurrence approach, and adaptive approach.

The naive approach for creating aggregated data for a user's bookmark collection is to count the occurrence of tags separately. If two tags are used in combination (co-occurred tags) by a certain user for annotating a certain bookmark, there is some kind

of semantic relationship between them. The co-occurrence approach is to calculate the weight of each pair of co-occurred tags for constructing a user profile. For a user, added bookmarks recently are more interested than old bookmarks. It makes a difference if a user has used a certain tag one day or one year ago. The *Add-A-Tag* algorithm is the adaptive approach which takes bookmarked dates into account. This approach extended the co-occurrence approach with the evaporation technique. Each time the profile is updated with tags from a newly added bookmark, the weight of each pair of co-occurred tags is decreased slightly by removing a small percentage of its current value.

As the majority of users are observed to be interested in a wide range of topics from different domains, a user profile in the form of a single set of tags is definitely inadequate. Further, it is obvious that documents related to the same interest of a user would be tagged by similar tags. Based on this observation, C. A. Yeung *et al.* [18] proposed a method for constructing user profiles which involves constructing a network of documents out of a personomy, applying community-discovery algorithms to divide the nodes into clusters, and extracting sets of tags which act as signatures of the clusters to reflect the interests of the users.

D. Zeng *et al.* [34] compared tagging user-based and traditional user-based collaborative filtering algorithm with a baseline, top-N algorithm on web page recommendation. In tag-based user profile, they considered tagging informations part of user profile. The profile of a user is a vector recording the frequency of tags ever used by the user. From their experimental results, the top-N method was lower than tagging and traditional user-based algorithm in almost all experimental data sets. In addition,

tag-based algorithm improved the precision by more than 10% over the traditional one. Similar trends were observed for recall for all datasets.

The results indicated that under the user-based recommendation framework, tags can be fruitfully exploited as they facilitate better user similarity calculation and help reduce sparsity related to past user-web page interactions.

In [4], M. J. Carman *et al.* discussed various models for generating a user profile using the information available in social bookmarking data for personalizing information retrieval.

There were five different models of tag-based profiles they proposed. Otherwise, they also developed content-based profiles which also had five models with the same approaches as tag-based profiles.

The *simple* profile is the same as the naive approach proposed by E. Michlmayr [20], which counts the occurrences of each distinct tag in a user's bookmark collection. The *common* profile ignores the tags in the bookmark without overlapping other tags in other bookmarks of a user. The *recent* profile considers the last k bookmarks only. The *decaying* profile, which weights older bookmarks less by multiplying their tag counts by a discounting factor, and E. Michlmayr's [20] *Add-A-Tag* algorithm are alike.

2.4 Common Similarity Measures

2.4.1 Jaccard Coefficient

Nowadays there are some common approaches for measuring similarity between profiles. The Jaccard coefficient, also known as the Jaccard index, is a statistic used for comparing the similarity and diversity of two sets. It measures the size of the intersection divided by the size of the union of two sets. We can treat an user profile as a set and each resource in the user's resource collection as an attribute which belongs to the set. We can treat each user's tag as an attribute from the user's tag-based profile similarly. The similarity measure of Jaccard coefficient is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (2.3)$$

where A and B are users' resource collections. When A and B are disjoint, the similarity is the lowest value 0. When A and B are the same set, the similarity is the highest value 1.

2.4.2 Cosine Similarity

Although the Jaccard coefficient can measure the similarity based on the overlap of the two sets, it ignores the weights of the attributes. Cosine similarity [28] is a similarity measure by computing the cosine of the angle between two vectors in n -dimensional space, and each dimension is one attribute with the associated weight. We define cosine similarity as:

$$\text{Cos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}. \quad (2.4)$$

The resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

For text matching, the attribute vectors A and B are usually the term frequency or TF-IDF [28] vectors of the documents. For recommender systems [1], we can utilize cosine similarity to find out the “neighbors”, similar users, of a specific user is vital and prerequisite for recommending remarkable items, and then we can estimate ratings for the items that have not been seen by a user from the ratings on the items given by the similar users. The more similar a neighbor and the target user are, the more weight rating will carry from the neighbor. This approach is called user-based collaborative filtering algorithm, which is one of the most popular approach in the research area of recommender systems.

2.4.3 Adjusted Cosine Similarity

In contrast with user-based collaborative filtering, item-based collaborative filtering [1, 29] is another facet to predict a user’s rating on an item. From the perspective of the user/item ratings matrix, item-based collaborative filtering is computed along the columns of the matrix, i.e., each pair in the co-rated set corresponds to a different user. Computing similarity using basic cosine similarity in item-based case has one important drawback, which is the differences in rating scale between different users are not taken into account. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair. Formally, the

similarity between items i and j using this scheme is given by:

$$AdjCos(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}}, \quad (2.5)$$

where $R_{u,i}$ is the rating of user u on item i , \bar{R}_u is the average of user u 's ratings and U is the set of users both rated item i and j .

2.4.4 Correlation-based Similarity

In this case, similarity between two items i and j is measured by computing the Pearson's correlation coefficient between them. To make the correlation computation accurate we only consider the set of users rated both item i and j which denoted as U , then the correlation similarity is defined as:

$$PCC(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_j)^2}}, \quad (2.6)$$

where \bar{R}_i is the average rating of item i .



Chapter 3

Semantic Similarity Measure for Tag-based User Profiles

A personal profile consists of simple factual data to describe a person, such as the name, age, educational background, or interests. In the era of Web 2.0, more and more web sites provide many kinds of services to people and make profits. Among those services, searching and recommendation are two popular ones for users provided by web sites. The better search result or recommendation result is achieved, the more users stick on the web site. For reaching the goal, the result is what the user wants, and it should be related to the user's interests. For this reason, many researchers try to produce user profiles as accurate as possible.

In the following sections, we introduce the background knowledge first, our problem definition and proposed solution orderly.

3.1 Background

3.1.1 Tag-based User Profile

Tagging was popularized by web sites associated with Web 2.0 and is an important feature of many Web 2.0 services. People annotate a resource (e.g. blog post, bookmark, image, video) with a set of tags to help them retrieve the resource later. Therefore, each tag should have semantic relation to the annotated resource for the user, and we can know which facets of the resource the user interests in. We define annotations as:

Definition 1. Model of Annotation

We define a set of users as \mathbf{U} , a set of tags as \mathbf{T} , and the resources in a collection as \mathbf{D} . Given a user $u \in \mathbf{U}$ and a resource $d \in \mathbf{D}$, we define $\text{annotate}(d, u) = \mathbf{T}' \in \mathbf{T}$.

Researchers [10, 9, 20, 18, 34, 4] utilize tags from users to produce tag-based profiles for exposing users' preferences clearly. Because of our work is the extension of tag-based user profiles, we make the definitions of tag-based user profiles first.

Definition 2. Tag-based User Profile

According to the Definition of User Profiling and Equation 4.1 in Chia-Chuan Hung's Master Thesis [10], we define a user u 's **tag-based user profile** as:

$$\text{Profile}_T(u) = \{(t_i, w_{u,i}) \mid t_i \in T_u\}, w_{u,i} = \frac{tf(u, t_i)}{\sum_i tf(u, t_i)} \quad (3.1)$$

where $tf(u, t_i)$ is the number of times user u used the tag t_i to annotate resources.

3.1.2 Semantic Similarity

In linguistics, semantics is the subfield that is devoted to the study of meaning, as inherent at the levels of words, phrases, sentences, and texts. And the study of semantic similarity [26, 23, 15] between words has been a part of psychology, computational linguistics, natural language processing, and information retrieval for many years. Psychologists use semantic similarity to describe similar degree between words, sentences, or contexts, and semantic similarity has become one ubiquitous and important variable that is often used to explain psychological phenomena.

In our proposed solution, we utilize the approach from Li *et al.* [15] based on WordNet [22], the approach from Speer *et al.* [30] based on ConceptNet [8], and the approach from Sahami [27] based on snippets returned from Google search engine to measure semantic similarities between tags. Here we briefly introduce their methods each.

WordNet

In Sec. 2.2.2, we introduced the benchmark in Table 2.2 for comparing approaches for measuring semantic similarity based on WordNet. The approach from Li *et al.* [15] performs the best on the benchmark among all approaches we have studied, so we adopt their approach for calculating semantic similarity between tags (words).

The difference between the approach from Li *et al.* and other approaches is that they transferred information sources nonlinearly. They argued that all first-hand information sources need to be properly processed in defining a similarity measure. First-

hand information sources are infinite to some extent, for example, the information content would tend to infinity if the probability of concept approaches zero in corpus. On the other hand, humans compare word similarity with a finite interval between completely similar and nothing similar. Thus, the transformation from the infinite interval to a finite interval is intuitively nonlinear. Among all strategies proposed by Li *et al.*, the best one is edge counting method which combines the shortest path length and the depth of subsumer nonlinearly. The shortest path length is the minimum length of path connecting the two concepts (synsets) containing the two words.

We adopt the approach which gives the best performance against human ratings in Table 2.2. The formula for similarity measure is

$$SemSim_{WN}(t_i, t_j) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (3.2)$$

where $\alpha \geq 0$ and $\beta > 0$ are parameters scaling the weight of shortest path length and the depth of subsumer respectively. From their experiment based on the benchmark data set, the optimal parameters for Eq. 3.2 are: $\alpha = 0.2$, $\beta = 0.6$.

ConceptNet

As we introduced fundamental elements of ConceptNet in Sec. 2.2.3, features are descriptions of concepts that complete an assertion about them. For example, the assertion “a trunk is part of a car” applies the feature (PartOf, *car*) to the concept *trunk*, and also applies the feature (*trunk*, PartOf) to the concept *car*. Each concept can then be associated with a vector in the space of possible features. Each assertion in ConceptNet has an integer confidence score which is initially 1. This score is automatically increased

when multiple users enter the same assertion and decreased when users enter contradictory assertions. The degree of similarity between two concepts is the dot product between their rows in the concept/feature matrix. However, these dot products have very high dimensionality and are difficult to work with.

Therefore, Speer *et al.* introduced a technique, AnalogySpace [30], to facilitate reasoning over a large knowledge base of natural language assertions that represent common sense knowledge. They utilized truncated singular value decomposition (SVD) which projects all of the concepts from the space of features into a space with many fewer dimensions. It also projects features from a space of concepts into the same reduced-dimensional space.

As AnalogySpace is an orthogonal transformation of the original concept and feature spaces, it can be used to compute similarity between concepts or between features by computing their dot products rapidly. Because all assertions are contributed by volunteers, ConceptNet contains some untrue concepts, so they take the concepts which involve at least 4 assertions into account. The researchers also have developed *Divisi*¹, a Python library for reasoning by analogy and association over common sense knowledge, for utilizing the knowledge from ConceptNet handily. Consequently, we use *Divisi* for calculating semantic similarity between tags and define the similarity measure as:

$$SemSim_{CN}(t_i, t_j) = v, \quad (3.3)$$

where $v \in [0.0, 1.0]$ is returned from the API of *Divisi* given t_i and t_j .

¹<http://divisi.media.mit.edu>

Google Snippets

The method proposed by Sahami *et al.* [27] is to measure the similarity by utilizing short text snippets returned from web search engine. Let w represent the query word, then:

1. Issue w as a query to a web search engine.
2. Let $S(w)$ be the set of n retrieved snippets s_1, s_2, \dots, s_n .
3. Compute the TF-IDF term vector v_i for each snippet $s_i \in S(w)$.
4. Let $C(w)$ be the centroid of the vectors v_i after normalization:

$$C(w) = \frac{1}{n} \sum_{i=1}^n \frac{v_i}{\|v_i\|}$$

In Step 1 we use Google search engine and in Step 2 we assign $n = 50$, which means we select top 50 results of each query word. In Step 3, we use the scheme TF-IDF for weighting terms in snippets, where the weight $w_{i,j}$ associated with term t_i in the snippet s_j is defined to be:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right),$$

where $tf_{i,j}$ is the frequency of term t_i in the snippet s_j , N is the total number of documents in the corpus, and df_i is the total number of documents that contain t_i . We compute N and df_i using a sample of about 300,000 documents from the web. Apparently, other weighting schemes are possible, but we choose TF-IDF here since it is most common used in the area of Information Retrieval.

Finally, given two tags t_i and t_j , we can measure the similarity between $C(t_i)$ and $C(t_j)$ by cosine similarity defined as:

$$SemSim_{GS}(t_i, t_j) = \frac{C(t_i) \cdot C(t_j)}{\|C(t_i)\| \|C(t_j)\|}. \quad (3.4)$$

3.2 Semantic Similarity between Tag-based User Profiles

Given two users u_a and u_b and their tag-based profiles as defined in Definition 2 , the **Goal** is to measure the similarity between u_a and u_b which represents their similar degree with the **Conditions** listed below.

We first define the goal as:

$$Sim_{STP}(u_a, u_b) = v_u, v_u \in \mathbb{R} \text{ and } v_u \in [0.0, 1.0] \quad (3.5)$$

Furthermore, the similarity measure have to satisfy the following properties as the conditions:

Property 1. Semantic Monotonicity

Let u'_a and u'_b are equivalent to u_a and u_b respectively except both using one more same tag t_i , then:

$$Sim_{STP}(u'_a, u'_b) \geq Sim_{STP}(u_a, u_b)$$

Property 2. Semantic Consistency

Given two users u_a and u_b and their tag collections T_{u_a} and T_{u_b} , let u'_a is equivalent

to u_a except using one more tag t_i , then:

$$\begin{aligned} Sim_{STP}(u'_a, u_b) &\geq Sim_{STP}(u_a, u_b), \\ \text{if } \max SemSim(t_i, t_b) &\geq \max SemSim(t_a, t_b) \end{aligned}$$

where $t_a \in T_{u_a}$, $t_b \in T_{u_b}$, and *SemSim* is any semantic similarity measure between two tags as described in Sec 3.1.2.

3.3 Proposed Solution

3.3.1 Semantic Tag-based User Profile

According to the characteristic of tags, researchers collected the set of tags the user used to build up his user profile. Each tag represents the user's attribute or interest, and has its associated weight. The more the associated weight, the more the tag can reveal the user's interest. The associated weight can be calculated by the number of times the user used the tag, and the weight can be adjusted by when the tag used or other issues.

But one common problem among their works is that they considered each tag independently. It means that one tag represents one interest, and any tag is irrelevant to any other tag. This is unreasonable to make such assumption because people have the knowledge or common sense about words are not the same but they can have some similar degree. For example, *design* is different from *layout*, but from our cognition there exists association between them instead of they are totally independent.

We propose *tag concept* near human intuition improving original tag-based user profile for resolving the problem described above. We define it as:

Definition 3. Tag Concept

Given a tag $t \in \mathbf{T}$, we identify a set of tags associated with weights as **tag concept** based on tag t by the following formula:

$$TC(t) = \{(t_j, w_j) \mid t_j \text{ is semantically similar to } t\}. \quad (3.6)$$

We also define scalar multiplication of a tag concept as:

$$r \cdot TC(t) = \{(t_j, r \cdot w_j) \mid t_j \text{ is semantically similar to } t, r \in \mathbb{R}\}. \quad (3.7)$$

Based on any semantic similarity measure described in Sec. 3.1.2, we can determine a set of tags which are similar to tag t semantically. We use a tag concept derived from a root tag in place of the root tag in a tag-based user profile. After replacing each tag with the corresponding tag concept, the new profile which consists of a set of tag concepts is a semantic tag-based user profile. We define it as:

Definition 4. Semantic Tag-based User Profile

Based on Definition 2 and Definition 3, we define a user u 's **semantic tag-based user profile** as:

$$\text{Profile}_{STP}(u) = \{t_i, TC_u(t_i) \mid t_i \in T_u\} \quad (3.8)$$

where $TC_u(t_i) = w_{u,i} \cdot TC(t_i)$.

3.3.2 Similarity Measure for Semantic Tag-based User Profiles

Based on semantic tag-based user profiles described above, we propose a method for measuring similarity between semantic tag-based user profiles, which in turn allows for

revealing similar degree between users. Furthermore, we can identify highly similar users given a target user for recommendation or other useful applications.

Because of a semantic tag-based profile consists of a set of tag concepts, we first define the similarity measure for tag concepts as:

Definition 5. Similarity Measure for Tag Concepts

Given two tags t_i and t_j , where $t_i, t_j \in \mathbf{T}$, we define a metric $Sim_{TC}(t_i, t_j)$ which is a similarity measure between tag concepts constructed from t_i and t_j as:

$$Sim_{TC}(t_i, t_j) = \begin{cases} SemSim(t_i, t_j), & \text{if } SemSim(t_i, t_j) \text{ exists} \\ \frac{TC(t_i) \cdot TC(t_j)}{\|TC(t_i)\| \|TC(t_j)\|}, & \text{otherwise.} \end{cases} \quad (3.9)$$

If the adopted semantic similarity measure for tags cannot measure the similarity between tag t_i and t_j , we measure the similarity from the tag concepts built from tag t_i and t_j . Finally, we define the similarity measure for semantic tag-based user profiles as:

Definition 6. Similarity Measure for Semantic Tag-based User Profiles

Give user u_i 's and u_j 's semantic tag-based profiles, where $u_i, u_j \in \mathbf{U}$, we define a metric $Sim_{STP}(u_i, u_j)$ which is a similarity measure between u_i 's and u_j 's semantic tag-based profiles as:

$$Sim_{STP}(u_a, u_b) = \sum_i w_{u_a, i} \cdot \max_j (Sim_{TC}(t_i, t_j)), \quad (3.10)$$

where $t_i \in T_{u_a}, t_j \in T_{u_b}$.

Chapter 4

Methodology of Semantic Tag-Based User Profiles

An important aspect of user profiles is whether they can truly reflect the interests or expertise of the users. Although the ratings a user give to resources are good sources for generating a user profile, we may create a user profile more precisely if we have more informations from the user. Furthermore, we can create user profiles for those users without giving ratings to the resources but other user-generated contents.

As more and more social media websites emerge, tags become rich user-generated informations. Tagging is used for managing resources originally, but tags generated by a user are desirable for exposing the user's interests like ratings. Instead of generating a user profile from ratings the user gave to resources, a vector of tags with associated weights are used.

In the following sections, we first introduce the similarity measure for tag-based

profiles with its deficiency, and then our proposed semantic tag-based profiles with the similarity measure.

4.1 Similarity Measure for Tag-based User Profiles

From a tag-based user profile defined in Def. 2, we can know what a user interests in and the degrees of the user's preferences. But tag-based user profiles make an assumption that the set of tags in a profile are independent, and there is no relation between any pair of tags. When calculating the similarity between two tag-based user profiles, *cosine similarity* [28] is the most common adopted measure which finds the cosine of the angle between two vectors, and it is also often used to compare documents in information retrieval. We define cosine similarity for calculating the similarity between two tag-based profiles as:

$$Sim_T(u_a, u_b) = \frac{\text{Profile}_T(u_a) \cdot \text{Profile}_T(u_b)}{\|\text{Profile}_T(u_a)\| \|\text{Profile}_T(u_b)\|} \quad (4.1)$$

where the numerator is a dot product of two vectors which consist of tag-based user profiles and the denominator is the magnitude of one vector multiplied by the magnitude of the other vector.

In Eq. 4.1, each distinct tag is one dimension of a vector and the associated weight is the length of the corresponding dimension. Unfortunately, it has missing and unsuitable for calculating similarity between tag-based profiles using cosine similarity because relatedness between tags are ignored spontaneously. For example, if one user has the tag `design` only and another user has the tag `art` only. By Eq. 4.1, the similarity of the two users are zero because `design` and `art` are different. But in fact,

they are not independent and there has some *semantic relation* between the design concept and the art concept from our knowledge. Therefore, we proposed “semantic tag-based profile” with its similarity measure for solving the problem.

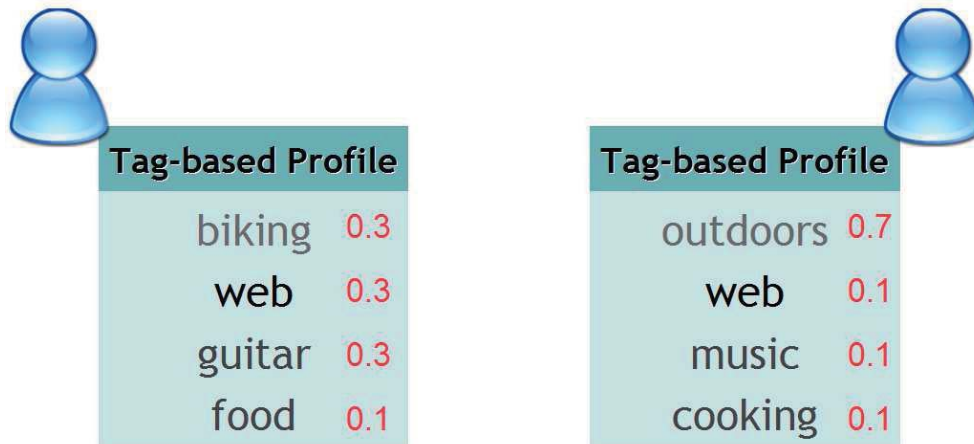


Figure 4.1: Using cosine similarity to measure the similarity between tag-based profiles is not suitable

4.2 Semantic Tag-based User Profile

By eliminating the deficiency on measuring similarity between two tag-based profiles presented by researchers [18, 34, 4, 20], we propose “semantic tag-based user profile”, which consists “tag concepts” constructed by spreading activation [5] and semantic similarity between tags derived from WordNet or ConceptNet.

Spreading activation is a method, which is based on supposed mechanisms of human memory operations, for searching associative networks or semantic networks. Originated from psychological studies, spreading activation was first introduced in

computer science in the area of artificial intelligence to provide a processing framework for semantic networks. Now it is adopted in many different areas such as cognitive science, psychology, databases and information retrieval.

The network data structure consists of nodes connected by edges. Nodes model objects or features of objects to be represented, and they are usually labelled with the name of the objects. Edges model relationships between nodes and they can be labelled and/or weighted. In our case nodes are tags and edges are undirected and weighted according to semantic similarity based on WordNet or ConceptNet.

The concept of spreading activation can be explained by a natural phenomenon. When we drop a stone in a pond, oscillation on surface transfers energy to neighborhood, and becomes smaller and smaller in amplitude due to water resistance. In this model, we can imagine each tag by a user as a stone. Its energy propagates from the most related tags to less relevant ones. A tag has an energy level indicating its relatedness to the primitive tag.

The processing technique is defined by a sequence of iterations, and each iteration is followed by another iteration until no new tag was marked as active in last iteration. In other words, each energy of activated tag in last iteration is not greater than the *firing threshold*. We define the steps of spreading activation as follows:

1. Initialize the graph setting all energies of nodes to zero and mark them as unactivated.
2. Set the root node to an initial energy $w = 1.0$ and mark it as active.
3. For each active node i in the graph:

4. For each edge e_{ij} connecting the active node i with the adjacent node j which is unactivated, add the spreading energy $w_i \cdot w_{ij} \cdot D$ to node j where w_{ij} is the weight of edge e_{ij} and D is the decay factor.
5. Mark all active nodes as activated.
6. If the nodes with augmented energy by Step 4 which is not greater than the firing threshold F , mark them as activated. If there exists the nodes with augmented energy which is greater than F , mark them as active and back to Step 3.

The decay factor D is like water resistance in the example above which controls the spreading energy. Usually D is set as 0.8. The firing threshold F is known as *activation constraint* which controls the spreading of the activation on the network. Moreover, it is possible to assign different threshold levels to each unit or set of units in relation to their meaning in the context of the application.

The procedure terminates when either there are no more nodes to mark as active or by *distance constraint*. Spreading activation should cease when it reaches nodes that are far away in terms of links covered to reach them from the root node. This corresponds to the simple heuristic rule that the strength of the relation between two nodes decreases with their semantic distance. Relations between two nodes directly connected are called first order relations. Relations between two nodes connected by means of an intermediate node are called second order relations, and so on. It is common to consider only first, second and, at most, third order relations.

For computing efficiency, we apply *fan-out constraint* to spreading activation. Spreading activation can cease at nodes with very high connectivity, that is at nodes connected

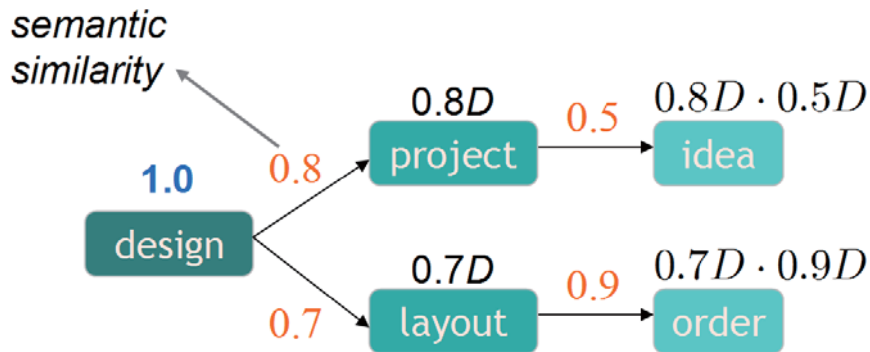


Figure 4.2: Construct a **tag concept** by spreading activation

to a very large number of other nodes, or spread energies to partial adjacent nodes only. The purpose of this constraint is to avoid a too wide spreading which could derive from nodes with a very broad semantic meaning and therefore connected to many other nodes.

We utilize spreading activation to construct a “tag concept” from a tag as in Fig. 4.2. From Sec. 3.1.2 we obtain semantic similarity between two tags using WordNet, ConceptNet, or Google snippets. If there exists semantic relation between two tags, semantic similarity is the weight of the edge connecting two tags. Otherwise, there has no edge between the two tags. In the graph, each tag is a node and edges represents semantic relations between tags as described above. For one tag t , we can obtain a set of tags which are marked as active once, and each tag in the set has associated weight which is the final energy acquired by the procedure of spreading activation. Eventually, we call the set of tags with their associate weights as “tag concept” and define it as:

$$TC(t_i) = \{(t_j, \phi(t_i, t_j)) \mid t_j \text{ is activated from } t_i\} \quad (4.2)$$

where tag t_i is the root tag of the tag concept, t_j is the tag activated by t_i and satisfied distance constraint with no more than third order relation, and $\phi(t_i, t_j)$ is the final energy or associated weight of tag t_j . Then we define scalar multiplication of tag concept as:

$$\begin{aligned} TC_u(t_i) &= w_{u,i} \cdot TC(t_i) \\ &= \{(t_j, w_{u,i} \cdot \phi(t_i, t_j)) \mid t_j \text{ is activated from } t_i, t_i \in T_u\} \end{aligned} \quad (4.3)$$

where $w_{u,i}$ is the associated weight of tag t_i defined in Eq. 3.1.

Each user has annotated his/her resource collection with many distinct tags. For each distinct tag, we set it as a root tag to construct a tag concept by the procedure of spreading activation described above. In conclusion, each user has the number of tag concepts which is the same as the number of distinct tags the user has used. And the user's *semantic tag-based user profile* consists of the set of tag concepts which defined as:

$$\text{Profile}_{ST}(u) = \{t_i, TC_u(t_i) \mid t_i \in T_u\}. \quad (4.4)$$

4.3 Similarity Measure for Smantic Tag-based User Profiles

From Eq. 4.1, tag-based user profile is not suitable for measuring similarity between two users with cosine similarity. For this reason, we propose an approach for mea-

asuring similarity between two semantic tag-based user profiles to reveal the similar degree.

Because semantic tag-based user profile is constructed by a number of tag concepts, firstly we need to present a method to calculate the similarity between tag concepts. A tag concept includes a set of activated tags with a root tag and their associated weights, and we can regard the tag concept as a vector which consists of the set of tags in the tag concept. The data structure of a tag concept is the same as the data structure of a tag-based profile, so we can use cosine similarity to find the cosine of the angle between two tag concepts defined in Eq. 3.9.

From Fig. 4.3, we show the advantage about using a tag concept instead of the root tag in the tag concept. Originally, the similarity between `driving` and `travel` is zero because they are different, and it is unreasonable. By applying cosine similarity, there exists the similarity between the two tag concepts because the tags, `driving` and `travel` and `trip` and `walking`, are overlapped. And it corresponds with what people think.

Maedche and Staab proposed an approach for measuring similarity between ontologies [17], which searches for the maximum overlap when comparing the two hierarchical structures. We take their approach as a reference to define the similarity measure for semantic tag-based user profiles in Eq. 3.10.

We preserve each maximum similarity among the pairs of one user's tag concept and the other user's all tag concepts as in Fig. 4.4, and then take the average of all maximum similarities as the similarity between two users. Another common formula is to calculate the average similarity of all the pairs mentioned above, but the value can

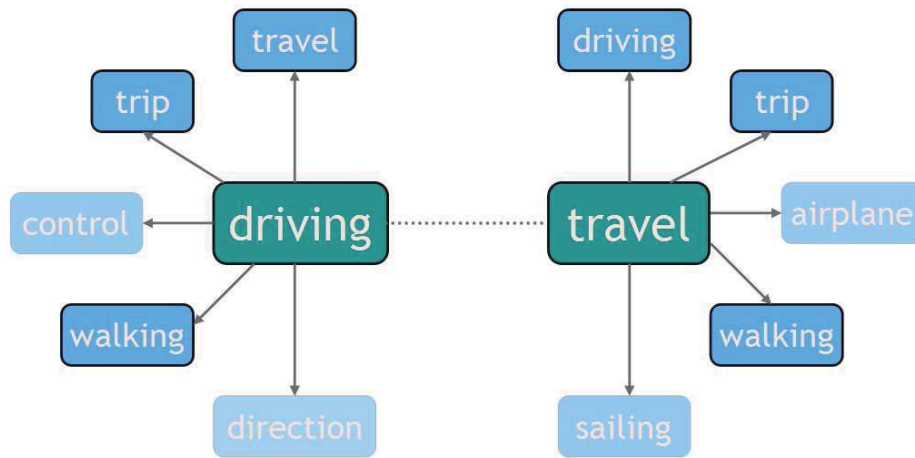


Figure 4.3: There exists semantic relation between the tag concept of driving and the tag concept of travel

be dropped enormously by adopting all tag concepts because users often have many divergent interests.

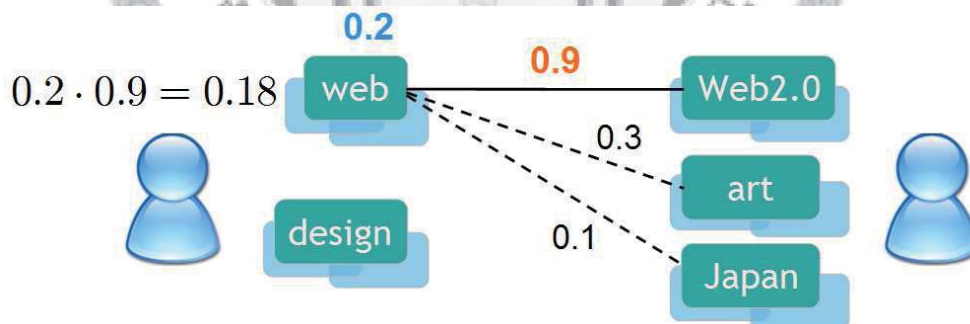


Figure 4.4: Select maximum similarity between among the pairs of one user's tag concept and the other user's all tag concepts

4.3.1 Property of the Similarity Measure

In the following we proof our proposed similarity measure for semantic tag-based profiles satisfying the properties described in Sec. 3.2.

Proof of Property 1: Semantic Monotonicity

Let user u 's total tag frequency $TTF(u) = \sum_j tf(u, t_j)$, $t_j \in T_u$, then:

(i) $t_i \notin T_{u_a}$:

After adding one more tag, the original weight $Sim_{ST}(u_a, u_b)$ is adjusted into $\frac{TTF(u_a)}{TTF(u_a)+1} Sim_{ST}(u_a, u_b)$, and then it appends the weight of the new tag concept $TC_{u_a}(t_i)$ which matches user u_b 's tag concept $TC_{u_b}(t_i)$.

$$\begin{aligned}
 & Sim_{ST}(u'_a, u'_b) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} Sim_{TC}(t_i, t_i) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} \\
 &\geq \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) \\
 &= Sim_{ST}(u_a, u_b)
 \end{aligned}$$

(ii) $t_i \in T_{u_a}$ but $t_i \notin T_{u_b}$:

Assume user u_a 's tag concept $TC_{u_a}(t_i)$ matched user u_b 's tag concept $TC_{u_b}(t_k)$ before using the tag t_i one more, we remove the weight of matching between them and then append the weight of matching between $TC_{u_a}(t_i)$ and $TC_{u_b}(t_i)$.

$$\begin{aligned}
 & Sim_{ST}(u'_a, u'_b) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} [Sim_{ST}(u_a, u_b) - \frac{tf(u_a, t_i)}{TTF(u_a)} Sim_{TC}(t_i, t_k)] + \frac{tf(u_a, t_i) + 1}{TTF(u_a) + 1} Sim_{TC}(t_i, t_i) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) - \frac{tf(u_a, t_i)}{TTF(u_a) + 1} Sim_{TC}(t_i, t_k) + \frac{tf(u_a, t_i) + 1}{TTF(u_a) + 1} \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} - \frac{tf(u_a, t_i)}{TTF(u_a) + 1} Sim_{TC}(t_i, t_k) + \frac{tf(u_a, t_i)}{TTF(u_a) + 1} \\
 &\geq \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{Sim_{ST}(u_a, u_b)}{TTF(u_a) + 1} - \frac{tf(u_a, t_i)}{TTF(u_a) + 1} + \frac{tf(u_a, t_i)}{TTF(u_a) + 1} \\
 &= Sim_{ST}(u_a, u_b)
 \end{aligned}$$

(iii) $t_i \in T_{u_a}$ and $t_i \in T_{u_b}$:

$$\begin{aligned}
 & Sim_{ST}(u'_a, u'_b) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} Sim_{TC}(t_i, t_i) \\
 &= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} \\
 &\geq \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} Sim_{TC}(u_a, u_b) \\
 &= Sim_{ST}(u_a, u_b)
 \end{aligned}$$

□

Proof of Property 2: Semantic Consistency

Based on the similarity measure for tag concepts in Definition 5:

$$\begin{aligned}
& Sim_{ST}(u'_a, u_b) \\
&= \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} \cdot \max Sim_{TC}(t_i, t_b) \\
&\geq \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} \cdot \max Sim_{TC}(t_a, t_b) \\
&\geq \frac{TTF(u_a)}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) + \frac{1}{TTF(u_a) + 1} Sim_{ST}(u_a, u_b) \\
&= Sim_{ST}(u_a, u_b)
\end{aligned}$$



□

Chapter 5

Experiment and Evaluation

In this section, we evaluate our proposed approach, semantic tag-based user profile formulated in Eq. 4.4, and the baseline approach, tag-based user profile formulated in Eq. 3.1, based on the data crawled from *Delicious* which is the most popular social bookmarking web service site.

5.1 Data Collection

For evaluation, we crawled the data including users and bookmarks. Each user on *Delicious* had an isolated web page for displaying the user's data as in Fig. 5.1, including his/her bookmark collection and the set of tags the user annotated on each bookmark. The data which we needed for evaluation includes the set of tags with the number of times each tag used by the user within his/her bookmark collection, and the set of tags which users annotated on a bookmark with the frequency of a tag annotated by

The screenshot shows a user's Deliculous profile page for 'mocat'. The page displays a list of bookmarks sorted by 'Most Recent'. Each bookmark entry includes the date, title, a count of other users who bookmarked it, and a set of tags. The tags are displayed as small buttons below each bookmark title.

Date	Title	Count	Tags
07 MAY 09	LaTeX Symbols	167	tex, latex, reference, symbol, math, thesis, research
06 MAY 09	JavaScript Visual Wordnet	140	visualization, wordnet, javascript, dictionary, graph, tools
22 MAR 09	Online LaTeX Equation Editor	185	tools, research, tex, web, math, equation
04 FEB 09	Photoshop Lady : Best Photoshop Tutorials Around the World	4600	reference, art, howto, tutorial, tips, webdesign, photoshop, design
22 JAN 09	Domain Evaluation - check your PageRank	10	tools, webservice, web, pagerank
07 JAN 09	delicious blog » Delicious is 5!	50	reference, research, del.icio.us, web2.0, bookmarking, blog
06 JAN 09	Can Social Bookmarking Improve Web Search?	112	research, tagging, web2.0, socialnetworking, social, search
04 JAN 09	[原文轉貼]如何寫論文：論文的起點 - MMDays	28	research, thesis

On the right side of the page, there is a 'Tags' section showing a list of tags and their counts. The top 10 tags are:

Tag	Count
research	30
web2.0	26
tagging	18
programming	15
socialnetworking	15
reference	12
tools	11
API	10
visualization	8
social	7

Below the top 10 tags, there are sections for 'Tag Bundles' and 'Programming' with their respective counts.

Figure 5.1: A user's data including his bookmark collection and used tags on Deliculous

users on the bookmark. In despite of Deliculous provided the API for retrieving the data, we cannot obtain all the data we need because it has some restrictions on using the API. Therefore, we crawled web pages directly and then parsed them with regular expression to retrieve what we want.

We crawled 39,459 users at first. From the statistic result, there were totally 9,149,239 distinct bookmarks among their bookmark collections. It would take a long time to crawl all bookmarks, and further, not all the data were satisfied the requirement of our evaluation. Accordingly, we filtered the bookmarks and then crawled the ones we needed.

5.2 Data Analysis

5.2.1 Data Filtering

Power-law distribution as Eq. 2.1 and Eq. 2.2 is a common observation of tagging systems. There are a few objects with high frequency and a lot of objects with low frequency in a power-law distribution. And it is important and useful for filtering the data based on power-law distribution.

We also observed the corresponding distribution among our crawled data. Each user has his/her bookmark collection, and the number of total bookmarks his/her owns. In Fig. 5.2, the vertical dimension is number of each user's total bookmarks and the horizontal dimension is the users ordered by the number of each user's total bookmarks. Both dimensions are logarithmic scale for representing power-law distribution explicitly. In our data the user with the most bookmarks has 56,663 bookmarks totally, the second rank user has 37,506 bookmarks, and the third rank user has 25,291 bookmarks. We can observe that the number of total bookmarks descends sharply and there are only 81 users whose own more than 10,000 bookmarks. On the other hand, 662 users have no bookmark, 632 users have one bookmark only, and 4,116 users have no greater than 10 bookmarks.

In order to construct a (semantic) tag-based profile effectively, we need the sufficient number of tags a user used for representing the user's interests, and it is the same as constructing profiles of bookmarks for evaluation later. From Golder and Huberman [6]'s experiment, the proportion of a used tag in a bookmark is nearly fixed after the first 100 users bookmarked the web page, so we keep the users and the bookmarks with

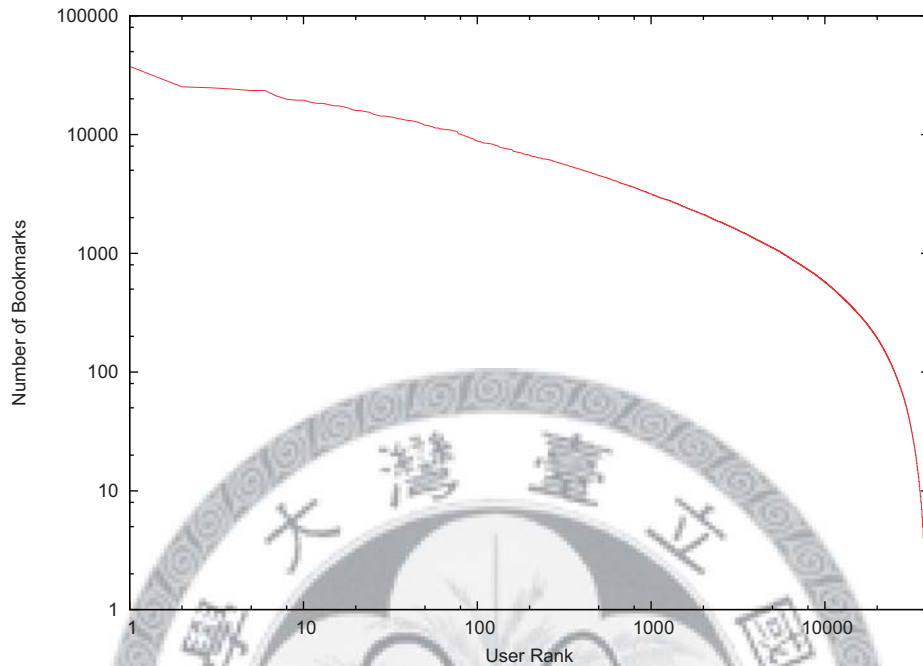


Figure 5.2: Distribution of the numbers of the users' total bookmarks

Data	User	Bookmark	Distinct Tag
Original	39,459	522,580 ^a	1,708,184
After filtering	20,578	80,000	1,353,828

^aEach bookmark is collected by at least 100 users.

Table 5.1: The crawled data for evaluation

more than 100 records. The number of the remaining bookmarks is 522,580 which is still a large number for our evaluation, so we adopt part of the remaining bookmarks which are collected by most evaluated users. Finally, we have 20,578 users and 80,000 bookmarks from the set of satisfied bookmarks for computing efficiency.

Rank	Tag	Frequency	Users	Rank	Tag	Frequency	Users
1	design	895,116	17,696	11	music	452,830	17,163
2	tools	762,283	17,178	12	howto	394,349	13,745
3	web2.0	701,278	16,072	13	css	394,261	14,429
4	software	667,346	17,056	14	google	387,546	16,984
5	blog	656,835	17,307	15	javascript	358,622	12,890
6	web	650,756	16,600	16	tutorial	340,763	14,532
7	webdesign	539,313	14,705	17	business	338,102	13,489
8	programming	514,125	13,543	18	free	326,727	13,922
9	video	505,929	17,967	19	development	322,871	12,009
10	reference	490,643	14,873	20	art	322,677	13,659

Table 5.2: The list of top 20 tags ordered by frequency with their frequencies and the numbers of users used them

5.2.2 Tag Coverages in Semantic Resources

After crawling the data from Delicious, the next step is to measure semantic similarities between distinct tags. But there are plenty of distinct tags in our crawled data set (and we listed top 20 tags in Table 5.1). If we measured all pairs of tags, it would take a long time to compute all semantic similarities. Besides, most tags are with low frequencies because the distribution of tag frequencies also fits the power-law distribution. For this reason, we select the top 15,000 tags and measure the semantic similarities of all pairs of the 15,000 tags.

Although there has about 150,000 words in WordNet and over 700,000 statements in ConceptNet, tags are freely chosen terms by users, including multi-language words, symbols, compound words, etc. The more tags are found in WordNet or ConceptNet, the richer semantic tag-based user profiles are constructed. Therefore, we check the coverage of tags in WordNet and ConceptNet and list the result in Fig. 5.3. The horizontal dimension is the top-n selected tags ordered by their frequencies, and the

vertical dimension is the proportion of the tags existing in WordNet or ConceptNet.

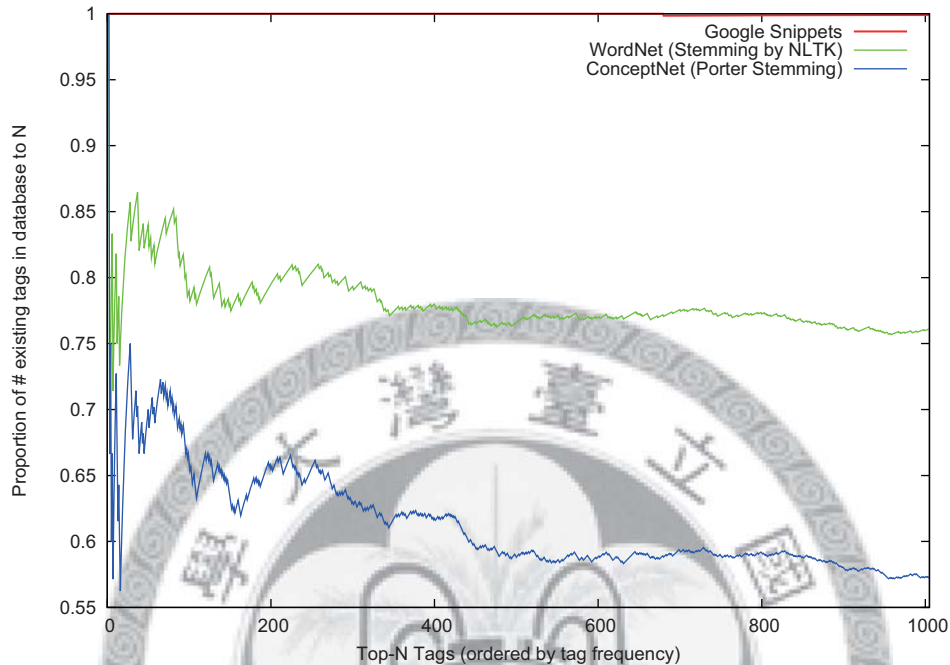


Figure 5.3: The Coverage of Tags in Semantic Resources

We utilize NLTK¹, which is a natural language toolkit for research and development with linguistic data, for checking a tag whether it is in the database of WordNet. From Fig. 5.3, we know the coverage of the tags in WordNet is higher than the coverage of the tags in ConceptNet with/without applying the Porter stemming algorithm[24]. The coverage of top 50 tags in WordNet to the total tags we selected is 84.0%, the coverage of top 100 and 300 tags are 79%, the coverage of top 500 tags is 76.6%, and the coverage of top 1,000 tags is still 76% high. The result shows that most tags with high frequencies exist in WordNet, and the coverage is almost stable from top 100 to top 1000 tags. The result from WordNet is acceptable, but the result from ConceptNet

¹<http://www.nltk.org>

is disappointing. The coverage of top 50 tags in ConceptNet is 32% only, the coverage of top 100 tags is 30%, and the coverage is 25% stably from top 300 to 1000 tags.

Since tags are freely chosen by users, we observed the tags in our crawled data have singular words and plural words, and nouns, adjectives, verbs, adverbs, etc. In order to improve the coverage of the tags, we use the Porter stemming algorithm for reducing inflected tags to their stem or root form. Stemming algorithms are common elements in query systems such as search engines for query expansion or indexing. For example, the words “fishing”, “fished”, and “fisher” are all reduced to the root word, “fish” by stemming algorithms. The Porter stemming algorithm is the most familiar stemming algorithm and it is also provided by NLTK, so we utilize it for stemming the tags which are not in ConceptNet, and check the root tags again.

With applying the Porter stemming algorithm, the coverage of top 50 tags in ConceptNet is up to 70%, the coverage of top 100 tags is 63%, and the coverage of top 500 and 1000 tags are about 57-58%. Although the result with stemming from ConceptNet is not as good as the result from WordNet, but it is much better than the result without stemming.

In addition, we manually divided some tags which are not in WordNet or ConceptNet to different kinds of tags as the following:

- Compound words: webdesign, toread, socialnetworking, opensource
- Technical words: web2.0, mysql, photoshop, skype
- Web sites: del.icio.us, youtube, twitter
- Abbreviations: hci, ui, api, apps

- Non-English words: 旅行, 部落格
- Non-words: !!!, #4, ***** , XD, 2008, 04/20.

5.2.3 Ratios of User's Tag Frequencies to Total Tag Frequency

After measuring semantic similarities between tags existed in WordNet or ConceptNet, we can construct the tag concepts from those tags by spreading activation. However, it takes a long time to construct users' all tag concepts due to most users used a lot of distinct tags. Therefore, we intend to select part of tags of each user to build tag concepts.

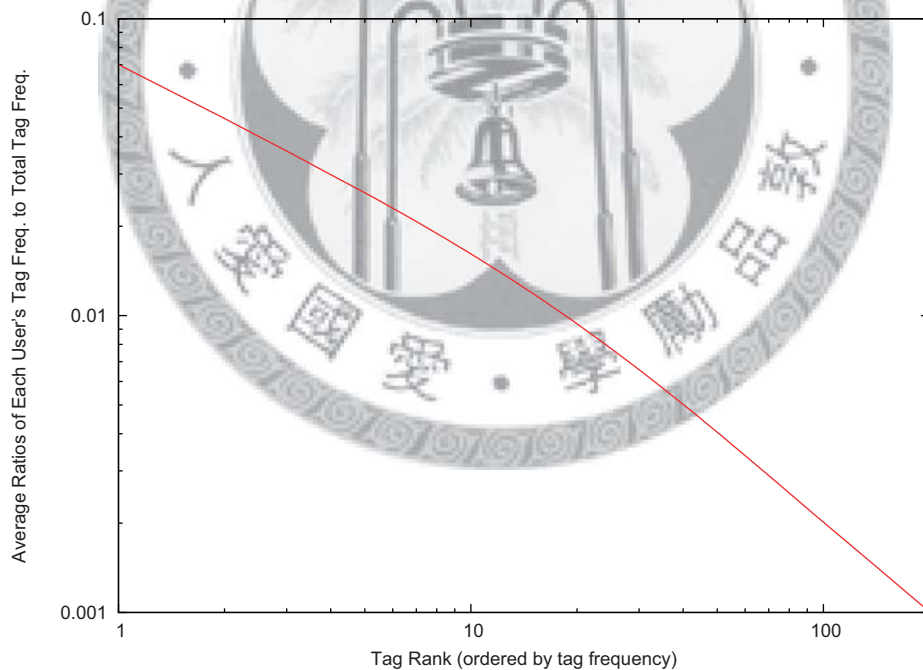


Figure 5.4: Average ratios of users' tag frequency of each rank to their total tag frequency

We calculated ratios of tag frequencies with different ranks to total tag frequency based on all users' data, and then calculated the average ratios of all users in our filtered data set. The result is showed in Fig. 5.4, and the distribution of the average tag ratios is a power-law distribution, which means part of tags dominate the weights of tags in a (semantic) tag-based user profile. Finally, We decide to build tag concepts from every user's top 30 tags for computing efficiency, and the average ratio of the rank 30th used tag is 0.6%. Another reason for selecting top 30 tags is that we can only retrieve top 30 tag frequencies of a bookmark within one request to *Delicious*. Because of we have to measure similarities between user profiles and bookmark profiles for empirical evaluation later, we also select users' top 30 tags only for fairness.

5.3 Example Result

We listed the result of semantic similarities between tag *design* and some other tags in Table 5.3. The result is from the relative semantic similarities based on WordNet, ConceptNet and Google snippets.

Tag	WordNet	ConceptNet	Google
web2.0	<i>None</i>	<i>None</i>	0.306
webdesign	<i>None</i>	<i>None</i>	0.682
designer	0.163	<i>None</i>	0.747
art	0.449	0.296	0.188
color	0.246	0.559	0.120
develop	0.519	0.016	0.254
happy	<i>None</i>	0.019	0.208
japan	0.245	0.085	0.032

Table 5.3: Example Result: Semantic Similarities between tag *design* and other tags

5.4 Empirical Evaluation

After crawling the data including the users and the bookmarks from Delicious and measuring the semantic similarities between top 15,000 tags based on WordNet, ConceptNet, and Google snippets, we can construct three semantic tag-based user profiles based on each semantic resource for a user to compare with the baseline method, tag-based user profile, described in Eq. 3.1.

We apply 5-fold cross validation to evaluate the performance of our proposed approaches. Cross validation is a technique for assessing how well the model you have learned from some training data is going to perform on future unseen data (or testing data). In 5-fold cross validation, every user's bookmark collection is partitioned into 5 subsets. The process is repeated 5 times. Each time a single subset is retained as the testing data, and the other 4 subsets are the training data. Finally, the evaluation result is from the average performance of 5 subsets as the testing set each. That is, for each user u 's bookmark collection D_u , we random select 80% bookmarks as the training data for constructing four user profiles including tag-based user profile, semantic tag-based user profile based on WordNet, ConceptNet, and Google snippets separately, and the other 20% bookmarks as the testing data known as the ground truths in our evaluation.

For each test of 5-fold cross validation, firstly, we construct three type of tag-based profiles for each bookmark, which consists of top 30 distinct tags with their associated weights, in the testing set. Secondly, for every user with one type of tag-based profile, we calculate the similarities between the user profile and the same type of bookmark

profiles. And then we sort the similarities to obtain the ranks of all the ground truth, the user's hidden bookmarks. The higher the ranks of the ground truth are, the more accurate the profile is. We can obtain three ranked lists for a user by three types of profiles totally, and we will show the evaluation results by different performance measures in the following subsections.

5.4.1 Precision-Recall Graph

In the area of Information Retrieval, the most common performance measures is *precision* and *recall* measures. Precision measure is the fraction of the bookmarks retrieved that are the ground truths, and recall measure is the proportion of the number of retrieved ground truths to the number of total ground truths. Precision and recall are measures for the entire testing set which do not account for the rankings of the ground truths in the retrieved data. In our evaluation, the higher the rankings of the ground truths, the better performance the profile reveals. Therefore, we consider the evaluation results by precision and recall measures at different cut-off points which are *precision at n* ($P@n$) and *recall at n* ($R@n$) listed below:

$$P(u)@n = \frac{|D_u \cap Q(u, n)|}{n} \quad (5.1)$$

$$R(u)@n = \frac{|D_u \cap Q(u, n)|}{|D_u|} \quad (5.2)$$

where $Q(u, n)$ are user u 's top n similar bookmarks among the testing set.

With the results from $P@n$ and $R@n$ measures at all cut-off points from 1 to the number of bookmarks in the testing set, we can plot a precision-recall graph, which shows the trade-off between precision and recall, as Fig. 5.5. Trying to increase recall

typically brings in more false data into the querying result, thereby reducing precision. Thus precision-recall graphs have a classical concave shape, which can depict the degradation of precision at n as one traverses the ranked list.

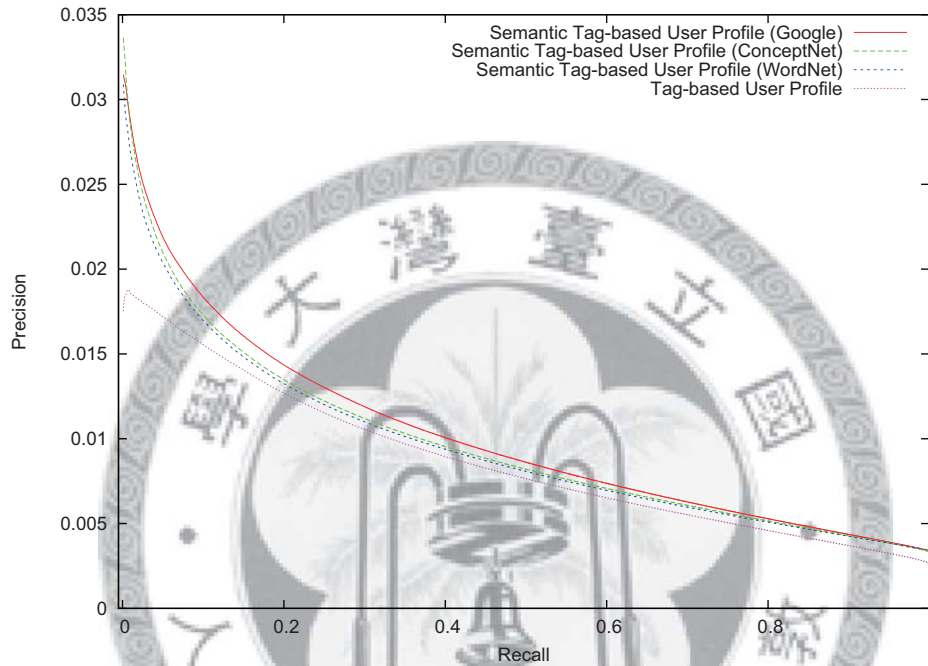


Figure 5.5: Evaluation Result by Precision-Recall Graph

The improvement for precision-recall graph is to increase both precision and recall. In other words, the entire curve must move up and out to the right so that both precision and recall are higher at every point along the curve. From the precision-recall graph in Fig. 5.5, the performances of three semantic tag-based user profiles are all better than the baseline, the tag-based user profile. The major differences between curves are within the range which recall value under 0.1, which means the ranks of a few top ground truths obtained by semantic tag-based user profiles outperform by tag-based

user profiles strongly.

5.4.2 Rank Accuracy Measures

Rank accuracy metrics measure the ability of a recommendation algorithm to produce a recommended ordering of items that matches how the user would have ordered the same items, and these metrics are more appropriate to evaluate algorithms that will be used to present ranked lists to the user. Thus we utilize two measures, mean reciprocal rank (MRR) and half-life utility measure [3], to compare the performances of three types of profiles.

The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer, and the mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries. We define $q_{u,i}$ as user u 's i -th similar ground truth and the formula of mean reciprocal rank as:

$$MRR = \frac{1}{|U|} \sum_i \frac{1}{rank(q_{i,1})} \quad (5.3)$$

where $rank(i)$ is a function for retrieving the rank of item i given a ranked list, and U is the set of users for evaluation.

From the result Fig. 5.6, we can see the performances of three semantic tag-based profiles (STBPs) are both better than the baseline, where the MRR from STBP based on WordNet is 0.093, the MRR from STBP based on ConceptNet is 0.098, the MRR from STBP based on Google snippets is 0.0975, and the MRR from the baseline is 0.067. The result shows the rank of each user's first similar ground truth in the testing

data by STBP is higher than the rank by tag-based profile (TBP) in a ranked list.

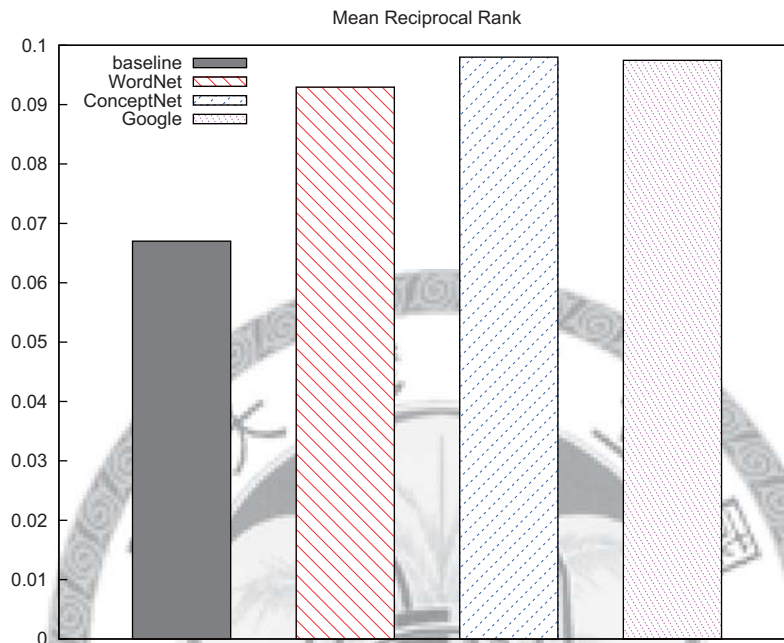


Figure 5.6: Evaluation Result by Mean Reciprocal Rank (MRR)

Mean reciprocal rank considers the rank of the first correct answer in a ranked list only. Moreover, we also should consider total ground truths in a ranked list. *Half-life utility metric* attempts to evaluate the utility of a ranked list, and the utility is defined as the difference between the user's rating for an item and the "default rating" for an item. The default rating is generally a neutral rating. Breese *et al.* [3] presented half-life utility metric for recommender systems that is designed for tasks where the user is presented with a ranked list of results, and is unlikely to browse very deeply into the list. For example, most Internet users will not browse very deeply into results returned by search engines.

In our data set, the rating of each bookmark is binary because a bookmark is whether in a user's bookmark collection or not, so we let the rating r be 1 if the bookmark is in the user's ground truth. We define the formula of the half-life utility metric as:

$$HU_u = \sum_i \frac{r}{2^{(\text{rank}(q_{u,i})-1)/(h-1)}} \quad (5.4)$$

where h is the half-life. The half-life is the rank of the item on the list such that there is a 50% chance that a user will view that item. We let h be 10 in Eq. 5.4.

The overall score for a data set across all users is shown in Eq. 5.5. HU_i^{max} is the maximum achievable utility if the system ranked the items in the exact order that user i ranked them. In other words, all user i 's hidden bookmarks are on the top of the ranked list.

$$HU = \frac{\sum_i HU_i}{\sum_i HU_i^{max}} \quad (5.5)$$

The result of half-life utility metric is shown in Fig. 5.7. The performances of two STBPs are also both better than TBP, where the utility from STBP based on WordNet is 0.0293, the utility based on ConceptNet is 0.0308, the utility based on Google snippets is 0.0313, and the utility from the baseline is 0.0244. From the half-life utility metric, we show semantic tag-based profiles are better than tag-based profiles by considering total ranks of the correct answers in a ranked list.

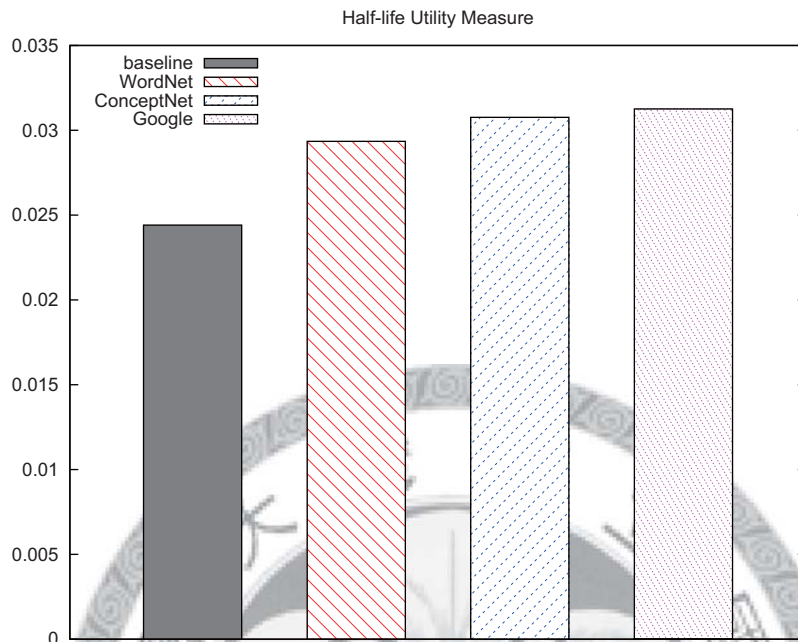


Figure 5.7: Evaluation Result by Half-life Utility Measure

5.5 User Study

Although we used the methods, precision-recall graph and mean reciprocal rank and half-life utility measure, to evaluate the performances of our proposed semantic tag-based profiles based on different semantic resources, this kind of evaluation considers users' history data only. All the unseen bookmarks are treated as wrong answers, and it is unreasonable to make this assumption. Therefore, we design a user study to recover the missing part of the empirical evaluation.

5.5.1 User Study Design

We design a web page as in Fig. 5.8 to collect the results from subjects. The requirements of a subject are the subject must have an account on *Delicious* with enough bookmarks for constructing tag-based profiles. For each subject, we construct three profiles from the subject's whole bookmark collection, including a semantic tag-based profile based on WordNet, a semantic tag-based profile based on ConceptNet, and a tag-based profile which is baseline, for evaluation. For each profile of a subject, we measure all similarities between the profile and the bookmarks in our data set excluding the bookmarks in the subject's collection. Then we select top 10 bookmarks from each profile and sort at most 30 bookmarks with a random order.

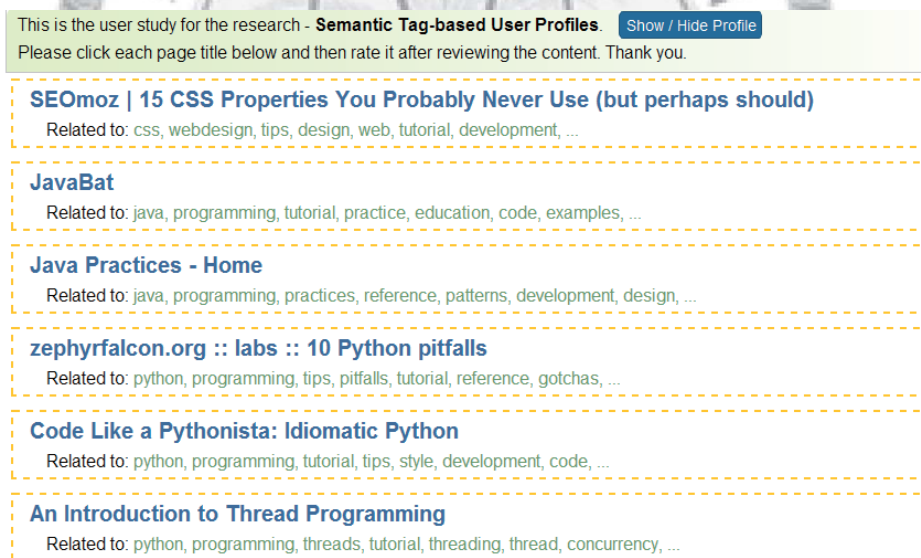


Figure 5.8: A Screen Shot of User Study

We put the data of selected bookmarks into the web page for clicking by subjects. Subjects are asked to click each item and give the rating after reading the page content.

They first can see the data of each item, including the page title and associated tags retrieved from *Delicious*. They also can see their profiles by tag cloud. After clicking the title of a item, the subject will see the information bar including the title, the rating stars and the text “More Info.” for showing the associated tags, and the page content displayed below as in Fig. 5.9. After reading the content of the item, the subject needs to give the rating according to his/her preference to the item. The range of the rating score is from 1 to 5. We also provide an icon for subjects to click if the server which holds the web page is error, the item is removed, the language of the text in the web page is unknown for subjects, etc.

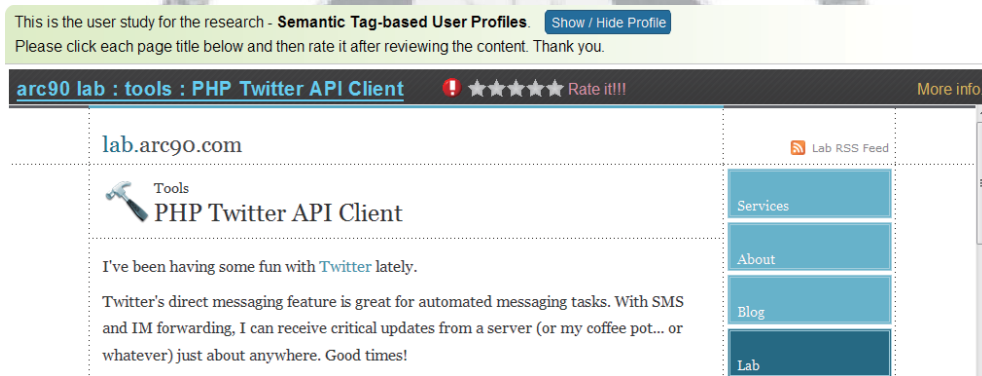


Figure 5.9: Let the subject give a rating after reading the web page content

5.5.2 User Study Result

We recruited 8 subjects for our user study, and they rated 211 web pages totally. We apply half-life utility measure to evaluate the performances of three different types of profiles. The rating r in half-life utility measure can be from 1 to 5 according to subjects' ratings, and the maximum achievable utility HU_i^{max} is gained by setting the

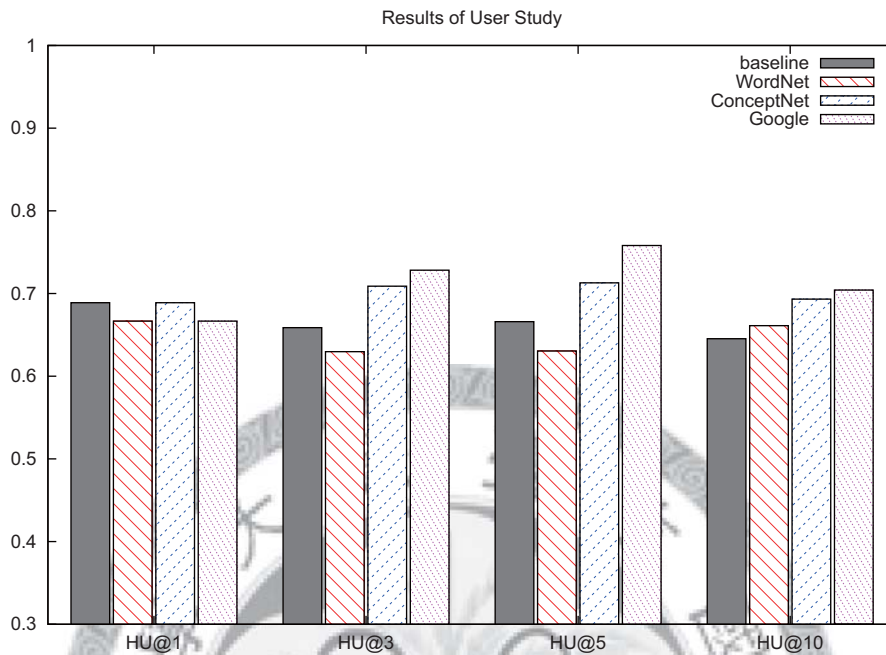


Figure 5.10: Evaluation Result of User Study

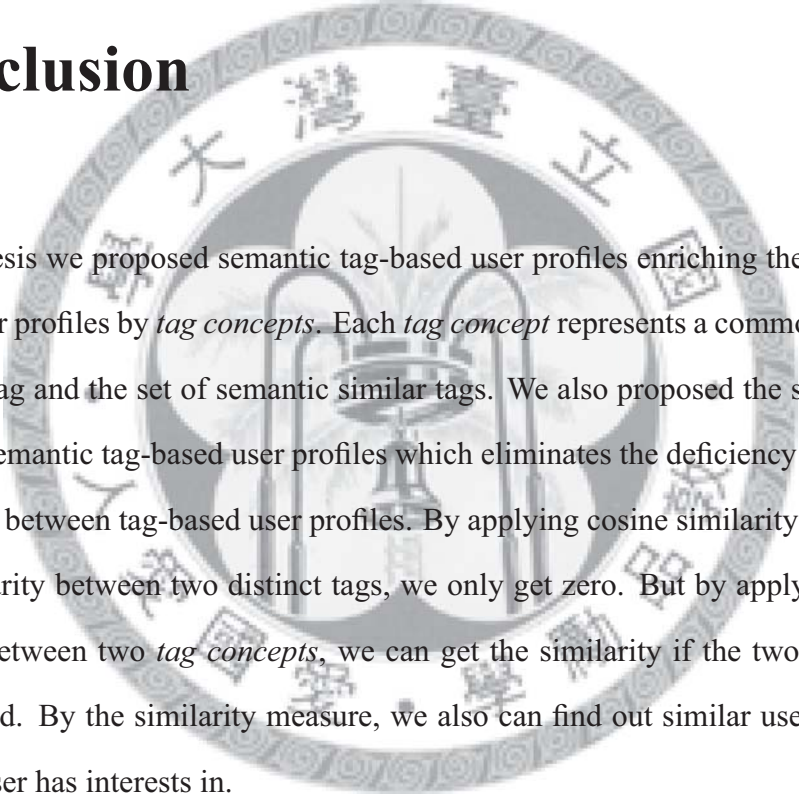
ratings of user i 's all items to 5.

We use $HU@n$ to view the average performances among all subjects' top- n item only, and we show the results including $HU@1$, $HU@3$, $HU@5$, and $HU@10$. From the results in Fig. 5.10, the most similar item measured by the baseline method showed the best performance, which means the subjects gave the ratings averagely higher than the top-1 items measured by semantic tag-based profiles. The utilities of semantic tag-based profiles based on ConceptNet are the best among $HU@3$, $HU@5$, and $HU@10$. The utilities of semantic tag-based profiles based on WordNet are a little lower than the utilities of the baseline method in $HU@3$ and $HU@5$, but it becomes better in $HU@10$.



Chapter 6

Conclusion



In this thesis we proposed semantic tag-based user profiles enriching the original tag-based user profiles by *tag concepts*. Each *tag concept* represents a common concept by the core tag and the set of semantic similar tags. We also proposed the similarity measure for semantic tag-based user profiles which eliminates the deficiency of measuring similarity between tag-based user profiles. By applying cosine similarity in measuring the similarity between two distinct tags, we only get zero. But by applying the same method between two *tag concepts*, we can get the similarity if the two concepts are overlapped. By the similarity measure, we also can find out similar users or identify items a user has interests in.

Based on a user's resource collection and associated sets of tags on social media sites, we could construct the semantic tag-based user profile containing the set of *tag concepts* to represent the user's interests. We introduced three semantic resources, WordNet and ConceptNet and Google snippets, with the associated approaches to mea-

sure semantic similarities between tags. We represented how to construct a *tag concept* from a tag by spreading activation with semantic similarities, and then we constructed a semantic tag-based user profile by a set of *tag concepts* from a user's resource collection with associated tags.

From empirical evaluation, we showed the performances of the semantic tag-based user profiles based on WordNet, ConceptNet, and Google snippets all were better than the performance of the tag-based user profile with the data set consisting 20,578 users and 80,000 bookmarks by 5-fold cross validation. From the result of user study, semantic tag-based user profiles based on ConceptNet show the best utility excluding considering top 1 only.

6.1 Summary of Contributions

- We proposed a semantic similarity measure for tag-based profiles with appropriate properties, and this measure eliminates the deficiency of measuring similarity by cosine similarity.
- We provided an insight into how the semantic tag-based profile of a user can be constructed from tags associated with the user's social media collection, and the semantic relations preserved in the profile could reflect the user's interests as the concepts.
- We proposed *tag concepts* capturing semantic relations between tags, and semantic similarities between tags could be measured based on different semantic

resources to represent different meanings. In this thesis we utilized WordNet, ConceptNet, and Google snippets to measure semantic similarity.

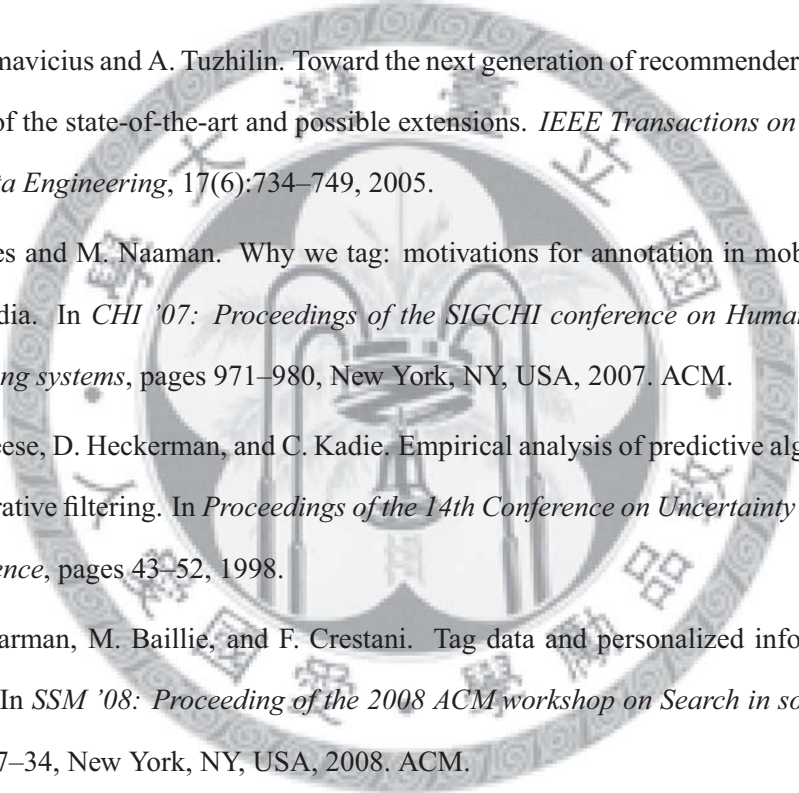
6.2 Future Work

According to our definition of semantic tag-based profiles, we can construct different profiles based on different approaches and semantic resources. However, it is possible to combine different semantic resources with associated similarity measures to construct one semantic tag-based profile revealing better performance. Based on the same tag with different semantic resources, we may construct tag concepts including distinct set of tags and associated weights. Thus, combining all tag concepts into one is an important issue to do in the future.

The problem about how to filter dissimilar tags in a tag concept is also a research issue. Further, if we can confirm dissimilar tags when measuring the similarity between tag concepts, we can obtain more accurate semantic similarity between tag concepts and between semantic tag-based profiles probably.

In this thesis, we construct semantic tag-base profiles based on tag-based profiles which tag weights are measured by a simple approach. However, tag weights can be determined by different approaches for different circumstances. For example, we can consider temporal factor and add more tag weights on the set of tags used recently. And we can combine those factors with our proposed solutions for different purposes.

Bibliography

- 
- [1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.
- [2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and on-line media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM.
- [3] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.
- [4] M. J. Carman, M. Baillie, and F. Crestani. Tag data and personalized information retrieval. In *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*, pages 27–34, New York, NY, USA, 2008. ACM.
- [5] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407 – 428, 1975.
- [6] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
- [7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In

- WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.
- [8] C. Havasi, R. Speer, and J. Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2007.
- [9] Y.-C. Huang. Tag-based profile presentation with semantic relationship, June 2008.
- [10] C.-C. Hung. Tag-based user profiling for social media recommendation, June 2008.
- [11] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *PKDD '07: Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514, Berlin, Heidelberg, 2007. Springer-Verlag.
- [12] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics*, pages 9008+, September 1997.
- [13] B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37–45, 1997.
- [14] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW '08: Proceedings of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM.
- [15] Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, 2003.
- [16] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

- [17] A. Maedche and S. Staab. Measuring similarity between ontologies. In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 251–263, London, UK, 2002. Springer-Verlag.
- [18] C. man Au Yeung, N. Gibbins, and N. Shadbolt. A study of user profile generation from folksonomies. In *Proceedings of the WWW 2008 Workshop on Social Web and Knowledge Management*, 2008.
- [19] D. L. Medin, R. L. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100:254–278, 1993.
- [20] E. Michlmayr and S. Cayzer. Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access. In *WWW '07: Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference*, May 2007.
- [21] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *ISWC '05: Proceedings of the 4th International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536, Galway, Ireland, 2005. Springer.
- [22] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.
- [23] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
- [24] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [25] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

- [26] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, 1965.
- [27] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA, 2006. ACM.
- [28] G. Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1983.
- [29] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.
- [30] R. Speer, C. Havasi, and H. Lieberman. Analogyspace: Reducing the dimensionality of common sense knowledge. In D. Fox and C. P. Gomes, editors, *AAAI '08: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 548–553. AAAI Press, 2008.
- [31] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.
- [32] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16, New York, NY, USA, 2005. ACM.
- [33] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1994. Association for Computational Linguistics.
- [34] D. Zeng and H. Li. How useful are tags? – an empirical analysis of collaborative tagging for web page recommendation. In *PAISI, PACCF and SOCO '08: Proceedings of the*

IEEE ISI 2008 PAISI, PACCF, and SOCO international workshops on Intelligence and Security Informatics, pages 320–330, Berlin, Heidelberg, 2008. Springer-Verlag.

