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

賦有語意關聯的視覺化標籤式使用者描述
Tag-based Profile Presentation with Semantic Relationship

黃怡靜
Yi-Ching Huang

指導教授： 許永真 博士
Advisor: Jane Yung-jen Hsu, Ph.D.

中華民國九十七年六月
June, 2008


國立臺灣大學
獎勵優秀教學

國立臺灣大學碩士學位論文
口試委員會審定書

賦有語意關聯的視覺化標籤式使用者描述

Tag-based Profile Presentation with Semantic Relationship

本論文係黃怡靜君（學號R95922045）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 97 年 6 月 4 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：許永真

（指導教授）

吳家麟　陳穎平

系主任　郭大維


國立臺灣大學
獎勵優秀教學

Acknowledgments

在碩士的兩年研究生涯，很幸運的進入台大資工所，認識了許多合作愉快的夥伴與師長，感謝他們的幫助、建議與鼓勵，讓我學習到研究的樂趣。這份碩士論文的順利完成，我要感謝所有陪伴我成長的人。

首先，要感謝我的指導教授：許永真教授，很高興能夠接受許教授的指導並與她合作，不僅讓我學到做研究時應有的積極理性的態度，更在我未來的人生方向上給了許多寶貴的引導，感謝許教授給我在研究上很高的自由度，讓我能夠自由發揮自己的想法，做自己有興趣的題目。感謝陳炳宇教授，在研究的過程中，給予我許多寶貴的建議與指導。另外，還要感謝我的口試委員吳家麟教授、林守德教授、陳穎平教授以及歐昱言教授，因為您們在口試時給予我的建議與指正，讓我發現許多盲點，使得論文得以順利完成。

感謝我這兩年的合作夥伴洪嘉涓，從碩一的微軟潛能創意盃競賽(Imagine Cup)，課堂上的專題合作，一起出國參加國際會議，到論文研究的合作。研究的過程中，一直有妳這位好夥伴一起共患難、共歡樂，在實驗室度過恐怖的停電，在家開著Skype一起做研究，在無數熬夜的日子裡，因為聽到對方喀喀的打字聲，讓在研究路上不再孤單，更加強了向前的鬥志。這段合作經驗真的相當愉快，很開心能夠認識這樣的一位好朋友。感謝我們的開心果陳麗徽，神經大條的程度讓人為之讚嘆，雖然每次都變成我們消遣的對象，但是妳的開朗總是影響整個實驗室，讓周遭總是充滿著歡笑。感謝實驗室的同學們，沈育仲、連家峻、吳祖佑、謝文芝、黃啟嘉和伍妮，很開心能與你們處在同一個實驗室

一起做研究、一起歡樂。感謝紀婉容學姐，總是耐心的聽我嘮叨，給我打氣的小餅乾，並且在我徬徨、遇到挫折時，給予我適時的建議與幫助，讓我順利克服困難。感謝同研究小組的學弟吳冠鋆、張琮傑，在研究上受到你們很多的幫助。感謝實驗室的學弟妹們，幫忙做實驗以及口試事宜的準備。感謝實驗室的所有人，跟你們相處真的是非常開心。

感謝我的家人，長久以來對我的呵護與包容，以及在碩士生涯的支持與鼓勵。感謝我的朋友們，總是帶給我歡樂，讓我抒發內心的壓力。感謝這一路上我身旁的所有人，因為有你們的陪伴，讓我的碩士生活更加多彩多姿。

Abstract

People construct *personal profiles* for self presentation and for obtaining online services. Profiles consisting of simple factual data provide an inadequate description of the individual, as they are often *incomplete*, mostly *subjective* and cannot reflect *dynamic changes*. This thesis explores the idea of ``you-are-what-you-tag'', namely, an individual can be effectively profiled by the tags associated with his/her social media. Specifically, this thesis proposes *semantic tag-based profiles*, profiles that can be represented as a set of semantically related and weighted tags. The strength of the semantic relationships between these tags are calculated using *common sense computing* and *co-occurrence* measurements. Moreover, different views of these profiles are visualized as tag clouds via a 3D switch effect. The proposed approach supports an intuitive and novel interface for people to browse/search through a social web site.

摘要

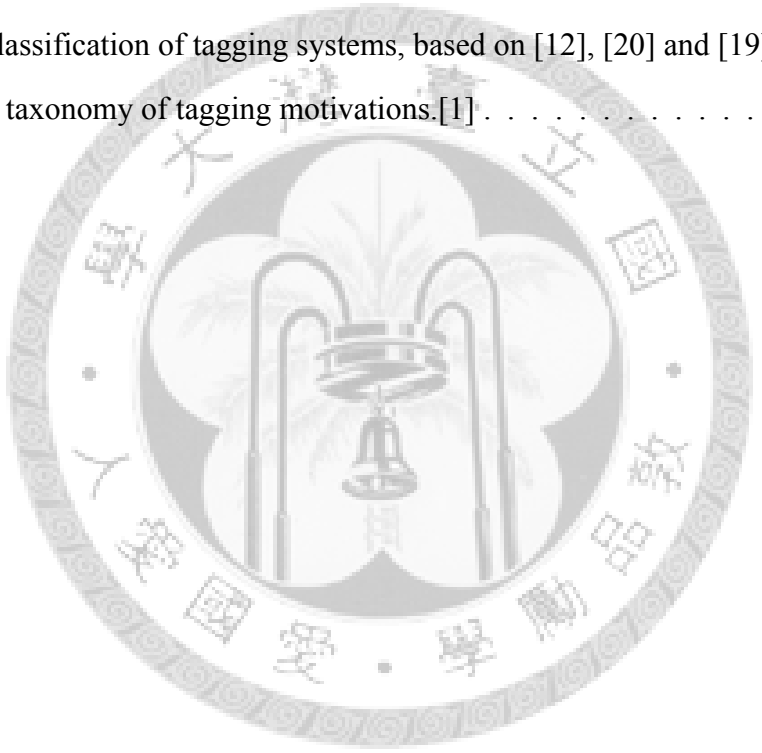
人們使用個人化的使用者描述（personal profile）來呈現自己的興趣與特色，並且取得許多線上的服務。這些使用者描述通常不夠完整，它們只包含了簡單的基本描述，而且只能主觀的呈現使用者自己的想法，無法反映出使用者興趣的動態變化，所以它們無法充分顯現個人的特質。在這篇論文裡，我們提出了附有語意關係的標籤式使用者描述的概念與方法。在概念上，意指我們可以利用所擁有的社群多媒體資料中所附含的標籤，有效的建立符合個人興趣與特質的使用者描述。在方法上，我們使用附有權重的標籤與不同強度的語意關係，來顯現使用者的想法與興趣。常識運算（common sense computing）與共現頻率（co-occurrence）能夠用來計算出不同標籤之間的語意關係。為了突顯出使用者標籤描述的特色與不同面向之間的差異，我們將使用者描述以標籤雲的方式做視覺化的呈現，並且加入了三維度的轉場效果，讓使用者更直覺、更自然的利用這樣的介面去搜尋在社群網路上的資料。


國立臺灣大學
獎勵優秀教學

Contents

List of Figures

List of Tables



Chapter 1

Introduction

1.1 Background

Online activity is becoming an increasingly important part of everyday life. People go online to look for jobs, keep in touch with friends and family, conduct business, talk about hobbies, and share their experience or feelings. In recent years, we see a great increase in the number of different kinds of social media web sites. The phenomenal rise of social media is transforming the average people from content readers to content publishers. Some popular social media services include del.icio.us[1] (social bookmarking), last.fm[2] (social music), flickr[3] (photo sharing), and YouTube[4] (video sharing), where people share a variety of media contents with their friends or the general public. Tagging is commonly used to add comments or descriptions about the media contents, or to help organize and retrieve relevant items.

Most social media sites supoprt some mechanism for tagging. For example, the

[1] http://del.icio.us
[2] http://www.last.fm/
[3] http://www.flickr.com/
[4] http://www.youtube.com/

bookmarks on del.icio.us may be tagged with the topics of interest to the user; a picture on Flickr may be tagged with its location, the event, the people and objects in the picture, the color or mood depicted in the picture. *Tagging* associates an object (e.g. a picture, a web page etc.) with a set of words, which represent the *semantic concepts* activated by the object at the cognitive level. Tagging provides a simple yet powerful way for organizing, retrieving and sharing different types of social media.

While categorization is a primarily subjective decision process, tagging is a social indexing process. In [18], Sinha succinctly pointed out that ``Tagging captures our individual conceptual associations, but does not force us to categorize. It enables loose coordination, but does not enforce the same interpretation of a concept. We could all tag items as `art' but mean very different things. That would create chaos in a shared folder scheme, but works well in a social tagging system." In addition, Sinha offered the following insightful oberservations.

- Tagging transforms web browsing from a *solitary* to a *social* experience. Tagging specific resources create ad-hoc groups, leading to ``wisdom of crowds".
- Tagging enables social coordination that is simultaneously more *direct* and *abstract* than collaborative filtering, as tags connect entities directly and enable tranfer of conceptual information.

1.2 Motivation

People construct personal profiles to present themselves and to obtain online services. A typical personal profile consisting of simple factual data, such as name, affiliation, or interests, provides an inadequate description of the individual. First of all, due to privacy concerns, most users are reluctant to provide more information than what is

required by the service. Secondly, these profiles are not specific because it cannot reveal the degree of user interest. Lastly, such simple user-specified profiles do not reflect dynamic changes, even though skills and interests of a person do evolve over time. Thus we plan to profile a user from his/her own media content on the Web and we propose an idea of tag-based profile.

An individual can be effectively profiled by the tags associated with his/her social media. Our basic assumption is that the rich online media produced/consumed by an individual can reveal important features about the person. Many online services today provide a platform for users to publish digital contents, which can be tagged. For example, the photo collections on Flickr show the people, the places, and the activities engaged by the user; the bookmarks on del.icio.us represent the topics of interest to the user; the blog posts on Blogger reflect the events, social interactions, or feelings experienced in the author's life. The user-specified tags associated with these personal collections of digital contents along with their comments provide meaningful descriptions of a person. However, utilizing a set of weighted tags as profile cannot describe the semantic relationship between tags. The semantic relationship between tags reveals much about the user's rich knowledge and interest.

1.3 Research Objectives

In the beginning of this thesis, we mentioned that user profile can represent user interests and is useful for some applications such as recommender systems. Others can also understand who you are via your user profile. Furthermore, personalized association and semantic relationship between interests can reveal user knowledge and preference. In this thesis, we propose semantic tag-based profiling to model user interest and knowledge more deeply. A set of weighted tags with semantic relationship can adap-

tively capture user interests over time. Common sense computing and co-occurrence measurement are defined to calculate the strength of semantic relationship between tags.

The primary research objective of this thesis is to present user profile in a visual way. Profile presentation can help others understand a user's interests and skills more easily and quickly. Thus we design a presentation tool to visualize the semantic tag-based profile.

1.4 Thesis Structure

The rest of this thesis is organized as follows. We start by going through some related research on tagging, folksonomy, common sense computing, user profiling and tag visualization in Chapter 2. The problem definition and the proposed solution for constructing semantic tag-based profile and for designing a profile visualization are explained in Chapter 3. In Chapter 4, we will define 3 types of knowledge and show the approach to calculating semantic relationships on tag-based profile in detail. Chapter 5 shows our design for presenting tag-based profile. Chapter 6 details the experiment for evaluating the semantic relationship. Finally, we make a conclusion in Chapter 7.

Chapter 2

Related Work

In this chapter, we present a brief survey of related work, including tagging, folksonomy, design of tagging systems, common sense computing, semantic similarity analysis, user profiling and tag visualization.

2.1 Tagging and Folksonomy

Tagging is a subjective indexing process of assigning freely chosen descriptive terms, also call tags, to digital media content for future navigation, filtering or search. In addition to organizing content, people can use tags to share their content with their families, friends and the general public.

Tag is a type of metadata (data about data) created by users. Traditionally metadata is created by either professionals or authors. In traditional approaches, end users are not involved in the creation of metadata and it is difficult to scale up with the vast amount of new contents being produced on the World Wide Web. In contrast to traditional approaches, tag is a more suitable approach for annotating digital media contents [13].

The social process in which users in various communities collaboratively tag publicly available resources and share contents is called ``*collaborative tagging*." In collaborative tagging systems, users share their tags for particular resources and a stable, community-wide pattern in tag usage emerges over time [6]. This pattern leads to an emergent, flat set of tags without a structured, hierarchical organization. This organization is called ``*folksonomy*," a user-generated classification, emerging through bottom-up consensus. It is a fusion of the words folks and taxonomy. The first use of the term folksonomy has been attributed to Thomas Vander Wal in 2004. Thomas defined folksonomy as the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. People use their own vocabulary to add explicit meanings to shared resources. The most value of a folksonomy is that it directly reflects the vocabulary of users. In our work, we try to extract a user's vocabulary or knowledge from his/her own media contents based on a combination of folksonomy and semantic analysis.

2.2 Design of Tagging Systems

In discussing tagging systems, two related issues are often overlooked. The first issue involves classification of tagging systems based on their design features; the second issue involves tagging incentives of users. Some previous studies for the two issues are introduced here.

In Stefaner's master thesis [19], he organizes the design features of tagging systems based on Marlow's classification [12] and a revised version presented in [20]. We follow Stefaner's organization of tagging systems by presenting the various dimensions of tagging systems in Table 2.1.

Incentives and motivations for users also play a significant role in affecting the tags

Table 2.1: Classification of tagging systems, based on [12], [20] and [19].

Dimension	Values	Explanation
Tagging Rights	Self-tagging	Users only can tag self-created resources
	Permission-based	Users can tag some resources
	Free-for-all	Users can tag all available resources
Source of Resources	User-generated content	Users tag self-generated content
	Provided content	Users tag content provided by the service
	External resources	Users tag resources not hosted by service
Resource Representation	Textual	Type of resource being tagged is textual
	Non-textual	Type of resource being tagged is non-textual (e.g. image or video)
Tagging Feedback	Blind	No awareness of community or own tags
	Viewable	Previously applied tags are presented
	Suggested	The system selects tag suggestions
Tag Aggregation	Set-model	Each distinct tag is only stored once
	Bag-model	Multiple applications of the same tag are counted
Vocabulary Control	Unrestricted vocabulary	Free–form annotation
	Managed vocabulary	Restricted vocabulary with regular updates
	Fixed vocabulary	Standardized classification
Vocabulary Connectivity	Unrelated tags	Keywords
	Associative	Authority file
	Hierarchical	Taxonomy/Classification
	Multi–hierarchical	Thesaurus/Faceted classification
	Typed	Ontology
Resource Connectivity	None	No specific relation between resources
	Links	Links between resources (e.g. web pages)
	Groups	Grouped resources (e.g. photo albums)
Automatic Tagging	None	Only user-defined tags
	Auto-tags	Automatically applied tags by resources analysis
	Automatic tag expansion	Automatically applied tags by user-defined tags

that emerge from collaborative tagging systems. Users are motivated both by personal needs and sociable interests. Marlow et al. categorized the motivations for tagging as *organizational* and *social*. The following list of incentives express the range of potential motivations that influence tagging behavior: (1) future retrieval; (2) contribution and sharing; (3) attract attention; (4) play and competition; (5) self presentation; (6) opinion expression. In [1], they extend Marlow et al.'s work and provide a more detailed taxonomy of tagging motivations on Flickr, as shown in Table 2.2. There are two dimensions: *sociality* and *function*. The first dimension, ``sociality,'' describes who uses the tags and uploads the photos, including friends/family and strangers. The second dimension, ``function'' refers to a tag's intended uses.

Table 2.2: A taxonomy of tagging motivations.[1]

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

2.3 Common Sense Computing

Simple descriptions are often used as tags to describe people's own contents. Choosing which tag for one content depends on people's preferences and knowledge. Tags are composed of words which have inherent semantic meanings in common sense. Tags can be analyzed with the help of common sense computing technology. Common sense knowledge collects a lot of human experience and encompasses knowledge about different aspects of typical everyday life. In this section, we introduce several popular

knowledge bases and explain how we use this computing technique briefly. Firstly, we will introduce two large-scale and general-purpose semantic knowledge bases, Cyc and WordNet. It costs most notable efforts to build them.

2.3.1 Cyc

The Cyc project begun in 1984 by Doug Lenat. Lenat's team tried to assemble a comprehensive ontology and database of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning. They used a logic framework to formalize common sense knowledge. Assertions are largely handcrafted by knowledge engineers at Cycorp, and as of 2003, Cyc has over 1.6 million facts interrelating more than 118000 concepts (source: cyc.com). The Cyc project has been described as "one of the most controversial endeavours of the artificial intelligence history,"[2] so it has inevitably some criticisms about the complexity of system, scalability problems, lack of any meaningful benchmark, etc. To use Cyc to reason about the text, it is necessary to understand its own language CycL. However, this mapping process is quite complex because all of the inherent ambiguity in natural language must be resolved to produce the unambiguous logical formulation required by CycL. The difficulty of applying Cyc to practical textual reasoning tasks, and the present unavailability of its full content to the general public, make it a prohibitive option for most textual-understanding tasks.

2.3.2 WordNet

WordNet [15][4] 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, general definitions, and records the various

semantic relations between these synonym sets. As of 2006, the database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs.

2.3.3 Open Mind Common Sense

Open Mind Common Sense (OMCS) is an artificial intelligence project, which is created by MIT Media Lab in 2000. It aims to construct a large common sense knowledge base from the general public. The collected data is contributed by web volunteers entering their common sense statements into the OMCS corpus. Since then they have gathered over 700,000 sentences of common sense knowledge from over 14,000 contributor from around the world, many with no special training in computer science. The OMCS corpus now consists of a tremendous range of different types of common sense knowledge, expressed in natural language.

2.3.4 ConceptNet

ConceptNet [11] is a an open-source tool for using the common sense knowledge collected in OMCS, developed by Liu and Singh. It is a semantic network with 20 relation-types that describe different relations among things, events, characters, etc. Figure 2.1 shows a concrete example of each relation-type from actual ConceptNet data.

2.3.5 Semantic Similarity Analysis

Measures of semantic similarity between concepts are widely used in Natural Language Processing and it refers to human judgments of the degree to which a given pair of concepts are related. In Pedersen et al.'s research [17], they develop a freely available

K-LINES (1.25 million assertions) (ConceptuallyRelatedTo 'bad breath' 'mint' '*f*=4;*i*=0;') (ThematicKLine 'wedding dress' 'veil' '*f*=9;*i*=0;') (SuperThematicKLine 'western civilisation' 'civilisation' '*f*=0;*i*=12;')
THINGS (52 000 assertions) (IsA 'horse' 'mammal' '*f*=17;*i*=3;') (PropertyOf 'fire' 'dangerous' '*f*=17;*i*=1;') (PartOf 'butterfly' 'wing' '*f*=5;*i*=1;') (MadeOf 'bacon' 'pig' '*f*=3;*i*=0;') (DefinedAs 'meat' 'flesh of animal' '*f*=2;*i*=1;')
AGENTS (104 000 assertions) (CapableOf 'dentist' 'pull tooth' '*f*=4;*i*=0;')
EVENTS (38 000 assertions) (PrerequisiteEventOf 'read letter' 'open envelope' '*f*=2;*i*=0;') (FirstSubeventOf 'start fire' 'light match' '*f*=2;*i*=3;') (SubeventOf 'play sport' 'score goal' '*f*=2;*i*=0;') (LastSubeventOf 'attend classical concert' 'applaud' '*f*=2;*i*=1;')
SPATIAL (36 000 assertions) (LocationOf 'army' 'in war' '*f*=3;*i*=0;')
CAUSAL (17 000 assertions) (EffectOf 'view video' 'entertainment' '*f*=2;*i*=0;') (DesirousEffectOf 'sweat' 'take shower' '*f*=3;*i*=1;')
FUNCTIONAL (115 000 assertions) (UsedFor 'fireplace' 'burn wood' '*f*=1;*i*=2;') (CapableOfReceivingAction 'drink' 'serve' '*f*=0;*i*=14;')
AFFECTIVE (34 000 assertions) (MotivationOf 'play game' 'compete' '*f*=3;*i*=0;') (DesireOf 'person' 'not be depressed' '*f*=2;*i*=0;')

Figure 2.1: Semantic Relation-types in ConceptNet

tool WordNet::Similarity, which provides six measures of similarity and three measures of relatedness between a pair of concepts (or word senses) based on the lexical database WordNet. A general classification of the measures and their relative advantage and disadvantage is provide in Fig 2.2.

Type	Name	Principle	Pro's	Con's
Path Finding	Path Length	Count of edges between concepts	- Simplicity	- Requires a rich and consistent hierarchy; - no multiple inheritance - WordNet nouns only - IS-A relations only
	Wu & Palmer	Path length to subsumer, scaled by subsumers path to root	- Simplicity	- WordNet nouns only - IS-A relations only
	Leacock & Chodorow	Finds the shortest path between concepts, and log smoothing	- Simplicity - Corrects for depth of hierarchy	- WordNet nouns only - IS-A relations only
	Hirst & St-Onge	Relies on synsets in WordNet	- Measures relatedness of all parts of speech - more than IS-A relations	- WordNet specific - Relies on synsets and relations not available in UMLS
Info. Content	Resnik	Information Content (IC) of the least common subsumer (LCS)	- Uses empirical information from corpora	- Does not use the IC of individual concepts, only that of the LCS - WordNet nouns only - IS-A relations only
	Jiang & Conrath; Lin	Extensions of Resnik; scale LCS by IC of concepts	-Accounts for the IC of individual concepts, only that of the LCS	- WordNet nouns only - IS-A relations only
Context Vector Measures	Patwardhan & Pedersen	Creates context vectors that represent the meaning of concepts derived from co-occurrence statistics of corpora	- Measures relatedness of all parts of speech - No underlying structure required - Uses empirical knowledge implicit in a corpus of data	- Definitions can be short, inconsistent - Computationally intensive

Figure 2.2: A classification of measures of semantic similarity and relatedness and their relative advantages and disadvantages.[16]

2.4 User Profiling

Research in [10] harvests profiles from social networking websites, such as Friendster[1], MySpace[2], and Orkut[3], to construct *InterestMap*, a network-style user profile to illustrate the relationship between interests and identities. Unlike traditional recommender systems, the proposed approach recommends by considering the interests of people instead of their historical behavior in a particular application. The InterestMap produces more accurate recommendations, and the preferences and interests of people in real life are modeled in an intuitive and visual fashion.

User profile can be provided by a user or can be built from his/her own content. In our work, we want to use tags on these content and the relationship between tags to represent a person's interest and characteristic. There is a similar idea in [14]. They construct user profiles from tagging data and they also compute the semantic relationship between tags using *co-occurrence*. A user profile is represented as tags and their relationships. They use a profile graph to represent a user, where nodes are tags used by this user and edges are the relations between tags and visualize a dynamic user profile by graph animation.

In contrast, Huang et al. [8] defined the personal, social and global views of user profiles from the tags associated with the social media content collected for the user. In addition, statistical and common sense reasoning were utilized to establish semantic connections among these tags.

[1] http://www.friendster.com
[2] http://www.myspace.com
[3] http://www.orkut.com

2.5 Tag-based and Social Visualization

On different social media websites, people use tags to describe their content and share with other people. Tagging is a social indexing process and contents can be categorized by any number of tags. As the number of tags increases, it becomes useful to view these tags visually. Therefore, there are more and more people getting involved in this issue about tag visualization. We present some types of tag visualization and social visualization. Firstly, we introduce the most popular visual representations: tag cloud and tag network. Next, we introduce some related work: tag orbitals and tag map. Finally, we introduce social network visualizations.

(a) Tag cloud on del.icio.us

(b) Tag cloud on flickr

Figure 2.3: Tag cloud on social media websites

2.5.1 Typical Tag Visualization

Tag Cloud

Tag cloud is the most common way to present tags. People use tags to organize their bookmarked URLs on del.icio.us and share their photos with others on Flickr (Fig 2.3). The tag cloud represents a set of tags as weighted lists. In general, people use tag frequency to determine which tags are more important than others and use font size and color to emphasize their importance. Typically, tag cloud is ordered alphabetically or by frequency. However, it is not easy to navigate when the number of tags increases day by day. In order to improve the tag cloud, some researchers try to cluster similar tags and show them together. In [7], they reduce the semantic density of a tag set and improve the visual consistency of the tag cloud layout. An approach to tag selection was proposed and a clustering algorithm was used to produce visual layout. Examples of their result are illustrated in Fig 2.4. Similar tags are placed together for easier navigation of pages by the users.

Tag Network/Graph

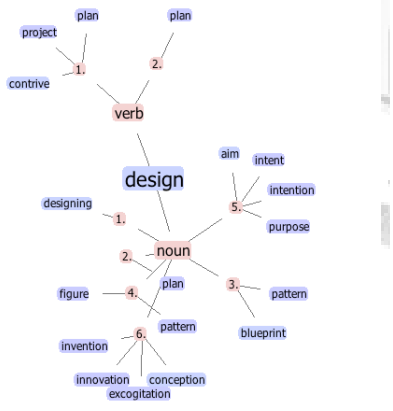
Tag network is usually used for presenting the relationships between tags. Through nodes and edges, people can realize the structure between tags. In Nearword[4], it shows word synonym based on the WordNet dictionary. People can use this visual tool to understand the different meanings of one word. Examples of Nearword are illustrated in Fig 2.5(a). In another work, they try to visualize tags via complex network diagrams. Given one specific tag, they will show related tags from del.icio.us (see Fig 2.5(b)). Their work presents the relationship between tags, but it is hard to interpret when the tag network is huge.

[4]http://www.intsysr.com/nearword.htm

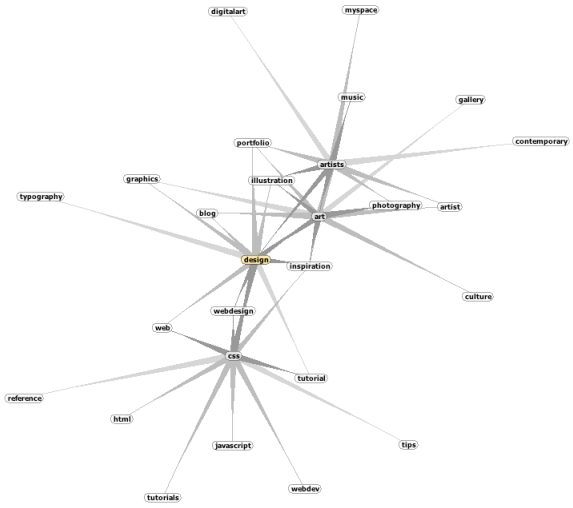
(a) current tag cloud on del.icio.us

(b) tag cloud with similar clustering

Figure 2.4: improve tagcloud with similar clustering



(a) Nearword visualization for ``*design*''



(b) GRAPH DEL.ICIO.US RELATED TAGS

Figure 2.5: Tag Network/Graph examples

2.5.2 Tag Orbital

TagOrbitals[9] is a tag visualization work designed by Bernard Kerr. In addition to tags and their inter-relationships, he included summary information about the tagged objects in his visualization (see Fig 2.6). The idea of TagOrbitals is based on the Bohr model of the atom. Each primary tag is composed of a series of concentric circles just like ``orbitals'' (see Fig 2.7(a)). The circle size is determined by tag weight. Each orbital level indicates the number of other tags used for each bookmark item (see Fig 2.7(b)). The first level show all tags which co-occur with the primary tag. The second level shows any set of two tags which co-occur with the primary tag, and so on.

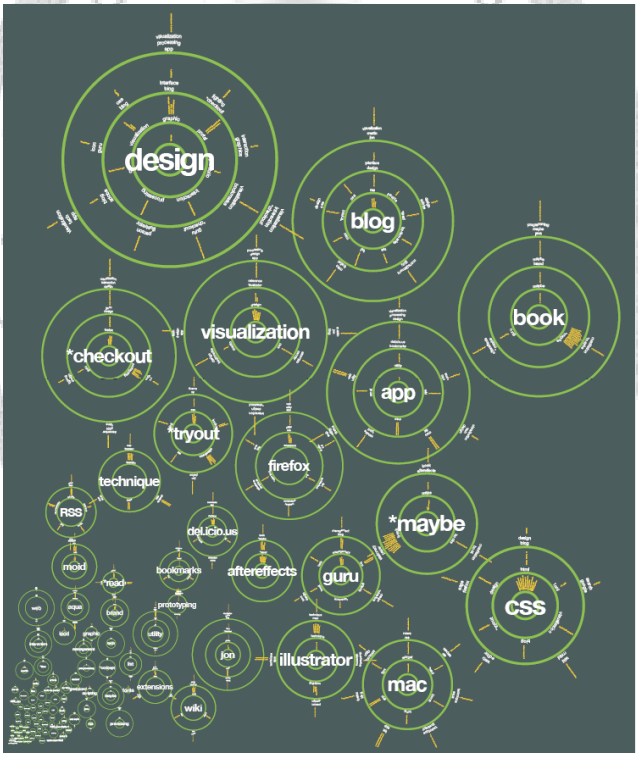


Figure 2.6: Tag Orbital

(a) related tags for design (b) url title for each tags

Figure 2.7: import tagcloud with similar clustering

Chapter 3

Tag-based Profile with Semantic Relationship

People describe interests and expertise in their profile and others can understand a person through these descriptions. User profile not only presents a user's interests but also is a basis for developing many applications such as recommender systems. However, such self-decribed profiles suffer from being incomplete, static, and not specific. In this sense, we focus on the problem of user profiling and modeling a user's knowledge. We want to model a user's knowledge and preferences into a cohesive user profile and design a profile presentation tool to show the user profile. We aim to let other people understand a person through such an advanced user profile.

Briefly speaking, the problem we want to solve is how to model a user's knowledge and preferences from tagging data and construct a tag-based user profile with semantic relationship. The Storyboard of semantic tag-based profile is shown in Figure 3.1. We focus on calculating the strength of semantic relationship on a tag-based profile.

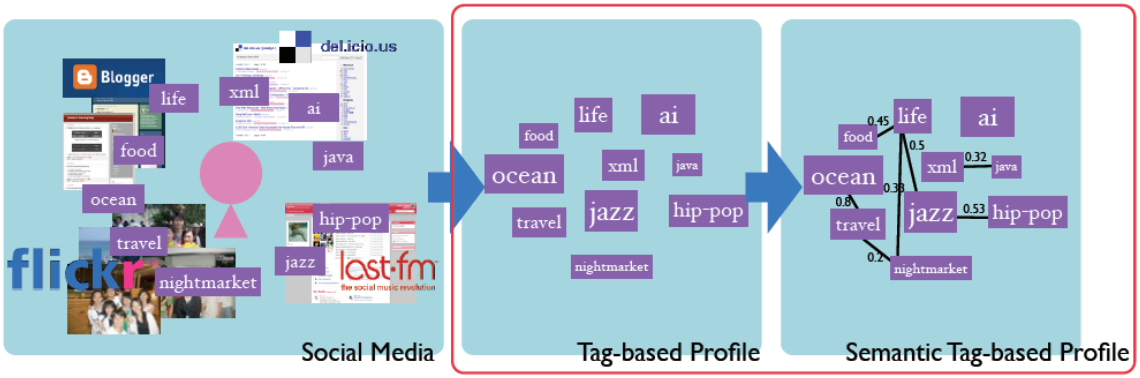


Figure 3.1: Storyboard

3.1 Problem Definition

We represent a tagging system by a tripartite graph with hyperedges.

Definition 1 (Model of Folksonomy) *The set of vertices is partitioned into three (possibly empty) disjoint sets* $A = \{a_1, ..., a_k\}$, $C = \{c_1, ..., c_l\}$, $I = \{i_1, ..., i_m\}$, *which respectively correspond to the set of actors (users), the set of concepts (tags) and the set of objects annotated (bookmarks, photos etc.). In this system, users tag objects with concepts, creating ternary associations amongst the user, the concept and the object. Thus the folksonomy is defined by a set of annotations* $T \subseteq A \times C \times I$. *We define the representing hypergraph of a folksonomy* T *as a (simple) tripartite hypergraph* $H(T) = \langle Vt, Et \rangle$ *where* $Vt = A \bigcup C \bigcup I$, $Et = \{\{a, c, i\} | (a, c, i) \in T\}$.

As mentioned in the previous section, we will focus on the problem of generating a semantic tag-based profile.

Definition 2 (Semantic Tag-based Profile) *A semantic tag-based profile* $Profile_{semantic} = (V, E, \Theta)$ *is formulated as an undirected graph* $G(V, E)$, *where* V *represents the node set and* E *represents the edge set in the graph. Each node* $v_i \in V$ *denotes one tag in*

the user's tag set. Each edge $e_{ij} \in E$ *represents the semantic relationship between tag* v_i *and tag* v_j *with the strength* $\theta_{ij} \in \Theta$.

Definition 3 (The Problem) *Given the set of user's tags* V*, the Problem is to find a method to compute the associated semantic tag-based profile* $Profile_{semantic} = (V, E, \Theta)$.

3.2 Proposed Solution

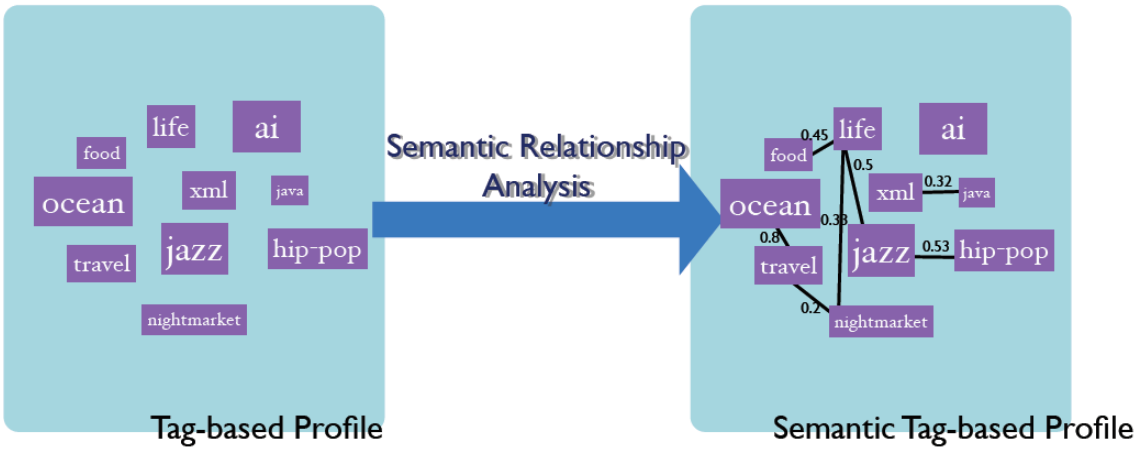


Figure 3.2: Proposed solution

3.2.1 Semantic Tag-based Profile

Each user is profiled as a set of weighted tags with semantic relationship from his/her media contents. Each relationship has its own strength. The profiling tags can be harvested from multiple data sources listed below:

- All data and descriptions in the registered user profile. This source of information

ranges from the bare minimal, e.g. only name and homepage URL for del.icio.us, to rich descriptions as in many social networking sites.

- Tags specified by the user for self description, or tags used explicitly by his/her friends to describe the given user.
- Tags associated with the user's collection of social media, which reflect his/her topics of interest as well as activities.

In [5], it shows that tag proportions for resources stabilize over time, which means that a set of weighted tags can represent a tag-based profile for a resource that does not change much once a sufficient number of tags have been collected. The benefit of using tags to profile people is that it not only preserves content characteristic but also captures personal preference. Therefore, we propose semantic tag-based profile to model users' knowledge and interests. Common sense computing and co-occurrence measurement are used to compute the strength of the relationships between tags. We believe that such a profile better fits a user's need and benefits more applications, particularly for recommender systems.

3.2.2 Tag-based Profile Presentation

Another contribution of this thesis is profile presentation. We design a visual tool to present the tag-based profile. Based on user's own documents, we can profile a user from three different views: personal, social and global. We construct the user profile from self-owned tag set (personal view) or from friends' tag sets (social view), or from all system users' tag sets (global view). We can understand a person from different viewpoints, through the opinions of others. Thus we design a 3D carousel effect to switch amongst the three profile views. We hope that users will like the presentation and show themselves to others.

Chapter 4

Semantic Relationship Analysis

Tagging has become a common practice for people to organize their own media contents for future navigation, filtering and searching (personal incentive). They use tagging to share their contents with their families, friends and others. These tags are freely chosen by users and reveal the interests and knowledge of the users. However, tags alone are not enough to portray a complete view of user interests and knowledge. we believe that semantic relationship between tags is also important for constructing a tag-based user profile. It reveals user knowledge in more detail. In this chapter, we define three types of knowledge: *personal association*, *community knowledge* and *global knowledge* and introduce how we extract these knowledge by co-occurrence and similarity methods.

Based on these three knowledge extracting methods, we propose a *semantic-based co-occurrence* method to calculate the strength of semantic relationships between tags. We want to both capture user preference and knowledge and construct a ``*semantic tag-based user profile*," which is a tag-based user profile with semantic relationship.

4.1 Three Types of Knowledge

An individual has unique knowledge and vocabulary to distinguish different objects in the world. Different people have different associations with the same object, but on the other hand, they also often share common knowledge about the same object. In a large community, users may share their own community knowledge and vocabulary with each other. In daily life, the general public share common sense knowledge to understand or recognize somebody, something or somewhere. Therefore, we define three types of knowledge: personal association, community vocabulary and global knowledge. They describe different characteristics of human knowledge and preference.

- **Personal Association:** The associations are different for different people. For example, if a user enjoys traveling in Japan, he/she would always associate ``travel" with ``japan. " This highly co-occuring relationship between ``travel" and ``japan" could reveal his/her personal preference on both. In our work, we use the *co-occurrence* method to calculate the joint probability of any two terms to determine those highly specific, personalized associations made by this user.

- **Community Knowledge:** Members of a community would share common knowledge with each other. For example, many programmers join an online club to discuss some RIA technologies like ``ajax" and ``flash." Programmers share their development experience and discuss some problems with others. These knowledge and vocabulary are understood by a group of people. In our work, we draw from the *folksonomy* phenomenon of a tagging system to get the popular tags on a particular resource (URL). Based on the ``wisdom of the crowd," these popular tags are representation of shared concepts on this resource under the shared vocabulary of this community.

- **Global Knowledge:** People use natural languages to communicate with others. These languages we use could be common sense in daily life and semantic meaning of words and texts from a thesaurus (or dictionary). In our work, we use both WordNet and ConceptNet to acquire knowledge from the general public. From WordNet, we can get the formal taxonomy definition of the English language, which also records the various semantic relations between words. ConceptNet, on the other hand, greatly expands the three semantic relations found in WordNet, to twenty, such as ``effect-of'', ``capable-of'', ``made-of'', etc. It contains practical knowledge from the general public and is useful for acquiring common sense knowledge.

4.2 Personal Association: Co-occurrence

In order to understand personal preference, we analyze personal bookmark data including URLs and tags. People bookmark URLs and tags them because they are interested in these topics or find the contents valuable to read again. This bookmark data could convey personal preference and more. Not only can we discover what a person likes but also understand how he/she thinks via his/her tagging history. We could take advantage of the relations between tags on his/her bookmark collection to better interpret his/her train of thoughts. To reach this goal, we propose a solution, ``co-occurrence,'' known as Jaccard coefficient. We assume that the more frequently two tags co-occur on the same documents(URLs), the more related the two tags are. Let A and B be the sets of documents described by two tags, co-occurrence is defined as Equation (4.1):

$$S_{co-occur}(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{4.1}$$

where $S_{co-occur}(A, B)$ is the co-occurrence of A and B. $|A \bigcap B|$ is the number of document in which tags co-occur and $|A \bigcup B|$ is the number of resources in which any one of the two tags occurs. In other words, we compute the proportion of tag overlap as tag similarity.

We apply the co-occurrence method on personal bookmark data. Figure 4.1 illustrates the idea of this method. We discover that problems may arise when calculating co-occurrence because user data may be sparse. The sparsity in user data arises from the tagging mechanism. Tagging is a free-style mechanism and people usually lose some tags on their collection for reasons such as being lazy or forgetful. Therefore, it is difficult to capture the personal preference on incomplete personal data. In order to solve this problem, we provide two methods to reinforce the ternary relations on users, tags, and URLs. In the next section, we introduce ``Social Wisdom," which could reinforce the relation between URLs and tags.

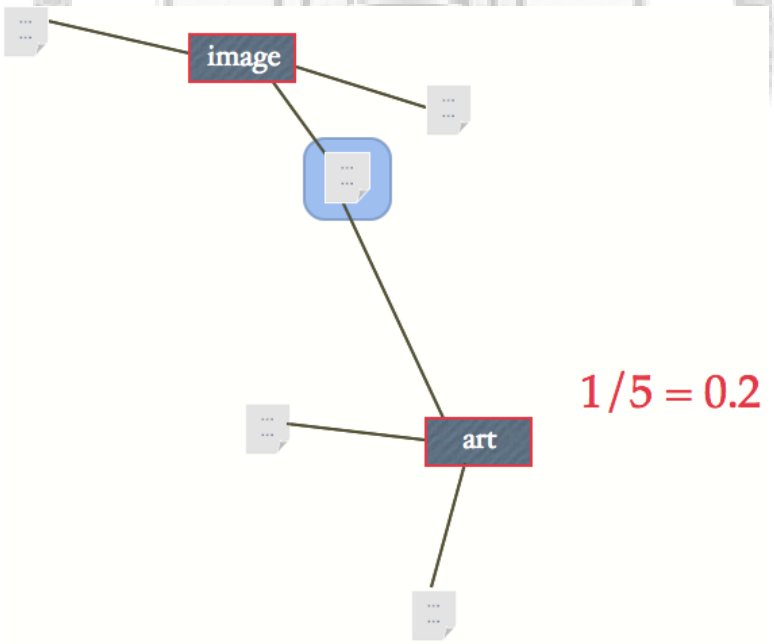


Figure 4.1: Co-occurrence method on personal bookmarking data.

4.3 Community Knowledge: Social Wisdom

In reality, user easily neglect some tags on URLs in various situations. For example, someone bookmarks a article about travel information in Taiwan. He assigns only one tag ``travelagent'' in a hurry and forgets to assign the relevant tag indicating the location ``Taiwan.'' (See Figure 4.2) This usually occurs in a collaborative system and some useful information may be lost. In order to enrich the number of tags on each URL, we utilize the ``wisdom of the crowd'' to add existing tags on URLs. These tags we add are from the user's past tagging history, as opposed to tags that the user never used, to better reflect his personal preference and to avoid incorrect results.

Figure 4.2: tag ``travelagent'' on real situation.

The purpose of ``social wisdom'' is to reinforce the links between tags and URLs on a user's bookmark collection. The equation of the social wisdom is defined as follows:

$$\begin{aligned} Tags(u_i) &= |Tags_{popular}(u_i) \bigcap Tags_{all}(p)| \\ SocialWisdom(t, u_i) &= AddLink(t, u_i), \forall t \in Tags(u_i) \end{aligned} \tag{4.2}$$

where $Tags_{popular}(u_i)$ refers to the top N popular tags for each URL u_i and $Tags_{all}(p)$ refers to all tags in the tag collection of user p. $AddLink(g, t, u_i)$ assigns the tag t on the URL u_i in the personal bookmark collection. The personal bookmark collection is show as a graph in Figure 4.3.

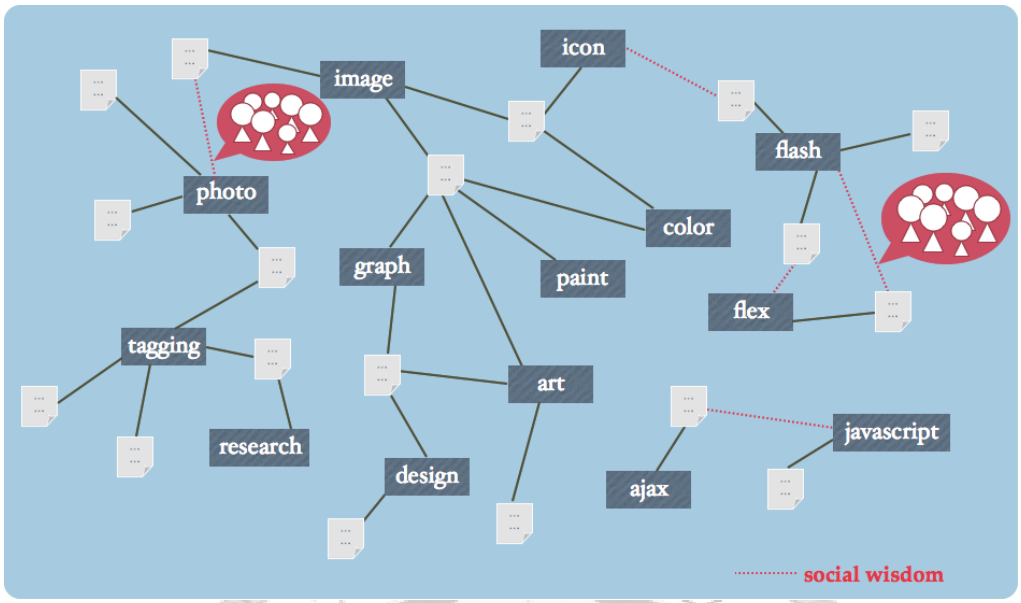


Figure 4.3: The personal tripartite graph with social wisdom

4.4 Global Knowledge: Semantic Similarity

In most cases, a tag is text with an inherent semantic meaning. People have commonly shared knowledge, known as common sense, on words used in daily life; moreover, some words have formal definitions in the dictionary, which is composed by professionals. We call these human knowledge, including common sense and formal definition, as global knowledge. In order to retrieve the global knowledge from tags, we establish the semantic similarity between tags by using two different kinds of databases, WordNet and ConceptNet.

4.4.1 WordNet-based similarity

WordNet is a semantic lexicon for the English language and it organizes nouns and verbs into hierarchies of is-a relations. We utilize WordNet::Similarity, which is a freely

available software package created by Pedersen et al. [17] to measure the semantic similarity of tags.

In this package, there are six measures of similarity, and three measures of relatedness. These measures are implemented as Perl modules which take as input two concepts and return a numeric value that represents the degree to which they are similar or related. In our work, we use a simple similarity measure ``path.'' It is a baseline that is equal to the inverse of the shortest path between concepts. Thus, we construct the WordNet-based semantic similarity on personal tag set by using this package.

4.4.2 ConceptNet-based similarity

In the previous section, we introduced how we measure the semantic similarity by WordNet. In this subsection, we introduce how to use common sense reasoning to obtain semantic similarity by ConceptNet, which is a freely available common sense knowledge base that provides a natural-language-processing toolkit for reasoning tasks including ``topic-jisting'', ``analogy-making'', and ``text summarization''.

ConceptNet is a semantic network created by Hugo Liu and Push Singh[11]. It collects common sense knowledge from the Open Mind Common Sense corpus and contains 300,000 nodes and 1.6 millions links, such as (IsA `apple' `red fruit') or (PropertyOf `game' `fun'). The ConceptNet toolkit provides node-level and document-level reasoning operations. Three functions on textual analysis[11] are introduced:

- GetContext(node): It accepts the input of a textual document which is then translated into a ConceptNet-compatible format. It finds the neighboring relevant concepts using spreading activation around this concept of the document. For example: the neighborhood of the concept ``music'' includes ``play violin'', ``play piano'', ``band'', etc.

- GuessConcept(node): It takes as input a document and a novel concept in that document, and it outputs a list of potential items which are analogous to the input concept. In other words, it can obtain analogous concepts from the concept of input document. For example: the concept of ``do exercise'' is analogous to ``ride bicycle'', ``play football'', etc.

- FindPathBetweenNodes(node1,node2) Find paths in the semantic network graph between two concepts.

Context of Concepts

Given two concepts a and b, the toolkit would determine all the concepts in the contextual neighborhood of a and b. We assume that C_a and C_b contain the contextual neighborhood concepts of a and b respectively. The similarity $S_c(a, b)$ between a and b based on context is defined as follows:

$$S_c(a, b) = \frac{|C_a \bigcap C_b|}{|C_a \bigcup C_b|} \tag{4.3}$$

where $|C_a \bigcap C_b|$ means the set of common concepts in C_a and C_b, and $|C_a \bigcup C_b|$ means the union set of C_a and C_b.

Analogous Concepts

Given two concepts a and b, the toolkit would determine all the analogous concepts of a and b. We assume that A_a and A_b respectively contain the analogous concepts of a and b. The similarity $S_a(a, b)$ between a and b based on analogous concepts is defined as follows:

$$S_a(a, b) = \frac{|A_a \bigcap A_b|}{|A_a \bigcup A_b|} \tag{4.4}$$

where $|A_a \bigcap C_b|$ means the set of common concepts in A_a and A_b. $|A_a \bigcup A_b|$ means the union set of A_a and A_b.

Number of paths between two concepts

Given two concepts a and b, the toolkit would determine all paths between a and b. We define that the path length between concepts a and b is the number of hops in each path. If there are more paths between two concepts, that means two concepts are more closed to each other. Thus, the similarity between them would be higher. For each path, the more hops between two concepts means they are farther away from each other; thus, the similarity would be lower. The path-based similarity is defined as follows:

$$S_p(a,b) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h} \tag{4.5}$$

where N is the total number of paths between a and b in the semantic network of ConceptNet and h_i means the number of hops in path i.

Combination of three measures

The final semantic similarity combines the three considerations: context, analogous concepts and number of paths. We compute it as a weighted sum of these measures. We use an equal weight on each measure and the ConceptNet-based semantic similarity is defined as follows:

$$CS(a,b) = W_c S_c(a,b) + W_a S_a(a,b) + W_p S_p(a,b) \tag{4.6}$$

where $W_c = W_a = W_p = 1/3$.

Having computed the ConceptNet-based semantic similarity between any two tags, the personal tripartite graph with semantic similarity is constructed and shown on Figure 4.4. In the next section, we will propose a semantic-based co-occurrence method

to calculate the semantic relationship between tags based on two personal semantic networks.

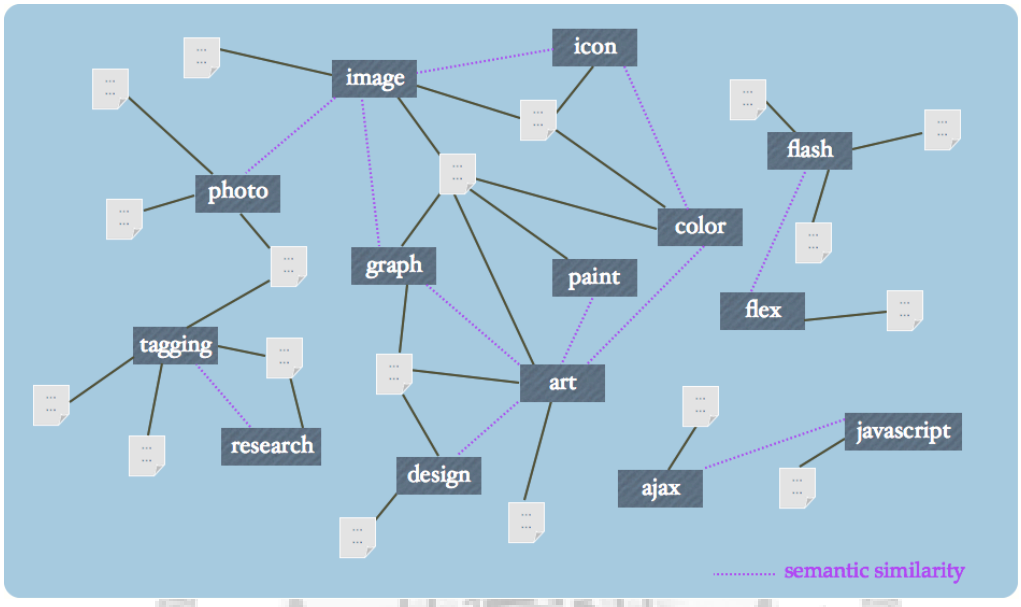


Figure 4.4: The personal tripartite graph with semantic similarity.

4.5 Semantic-based Co-occurrence

In this section, we introduce our method to calculate the semantic relationship between tags. Firstly, we propose an idea of ``Tag Concept'' and how to get tag concept based on semantic similarity. Next, we introduce how to calculate co-occurrence based on the tag concept. This method not only considers personalized association (co-occurrence), but also global knowledge (semantic similarity).

4.5.1 Tag Concept Based on Semantic Similarity

Spreading Activation

Concepts and ideas in the human brain have been shown to be semantically linked. Thus thinking about (or firing) one concept primes other related concepts, making them more likely to fire in the near future. In our work, we use the semantic network to model user knowledge and to find personalized associations.

We use a spreading activation algorithm [3] to conduct inferences and compute the similarity among tags. The input tag as a first node has highest level of energy and spreads a fraction of its energy to relevant tags. The value of spreading energy is directly proportional to the weight between tags. The energy of any tag after a spreading step is calculated by Equation (4.7):

$$Energy(t_j) = \sum_{i=inDegree(t_j)} Energy(t_i) * Weight(t_i, t_j) * \alpha \tag{4.7}$$

where t_j is the activation level of tag t_j, t_i is a tag connected to tag t_j, $Energy(t_j)$ is energy of t_j acquired from t_i, and $Weight(t_i, t_j)$ is a link weight between t_i and t_j. $inDegree(t_j)$ means the number of inlinks on tag t_j. If $Energy(t_j)$ exceeds a threshold f, tag t_j will be activated in next activation level. The energy of the tag would decrease at a ratio α step by step, and stop until no new tags are activated. Finally, we collect the activated tags which are the related tags.

Tag Concept

Applying spreading activation on personal knowledge network, we can identify related tags given any target tag. We define these related tags of the target tag together as ``tag concept'' of this target tag. In Figure 4.5, the tag concept of ``image'' is a set containing

``photo," ``graph," ``icon" and ``image" (target tag).

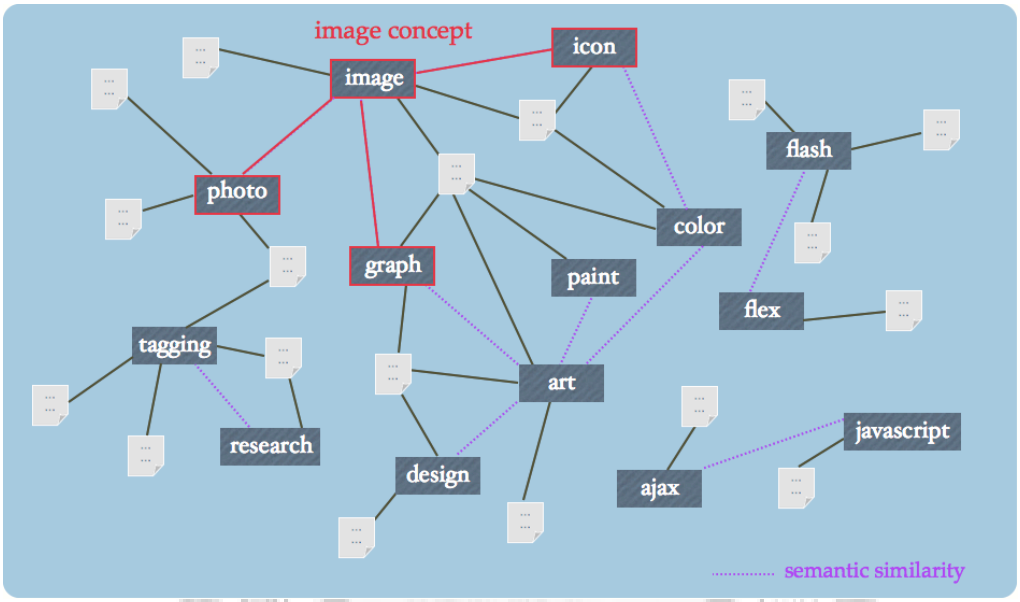


Figure 4.5: Tag concept of ``image" based on semantic similarity

4.5.2 Semantic Co-occurrence Based on Tag Concept

We propose a ``semantic co-occurrence" approach to calculating the semantic relationship between tags based on the tag concept. We calculate the number of co-occurring tag concepts on the same documents. The equation of semantic co-occurrence is defined by Equation 4.8.

$$S_{semantic}(a,b) = \frac{|TagConcept(a) \bigcap TagConcept(b)|}{|TagConcept(a) \bigcup TagConcept(b)|} \tag{4.8}$$

where $TagConcept(a)$ means the related tag set of target tag a by spreading activation and $S_{semantic}(a,b)$ is the co-occurrence of $TagConcept(a)$ and $TagConcept(b)$. The numerator $|TagConcept(a) \bigcap TagConcept(b)|$ is the number of documents in which the two tag concepts co-occur; the denominator $|TagConcept(a) \bigcup TagConcept(b)|$,

on the other hand, is the number of resources in which any one of the two tag concepts is present. In other words, we compute the proportion of overlapping tag concepts as tag semantic similarity. Figure 4.6 shows the idea of semantic co-occurrence.

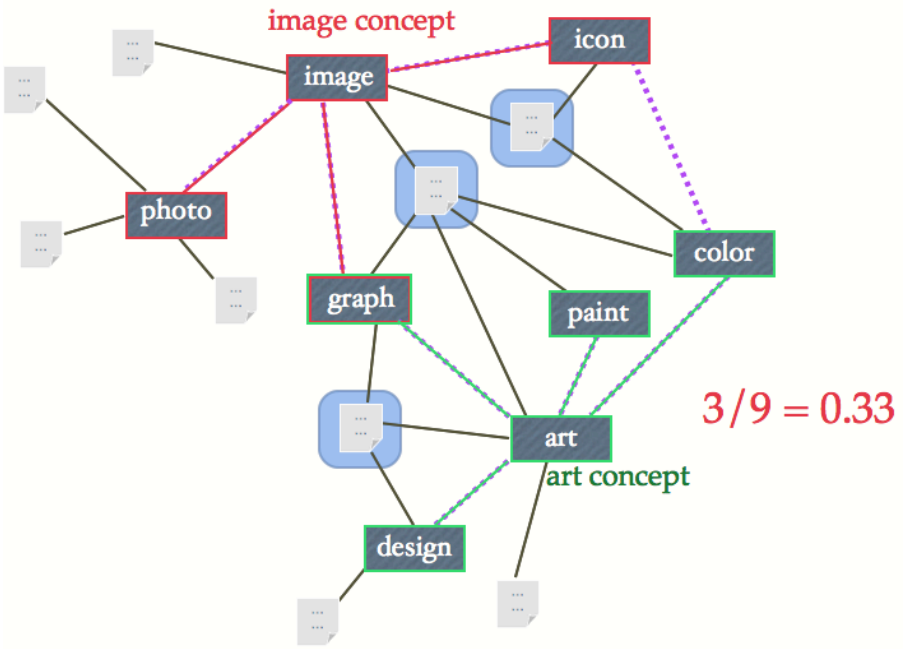


Figure 4.6: Semantic co-occurrence based on tag concept

Chapter 5

Tag-based Profile Presentation

In the previous chapter, we presented the creation of a semantic tag-based profile for the purpose of extracting user interests from personal media content. In this chapter, we propose the design of a visual tool to present the semantic tag-based profile.

5.1 Data Characteristic

The semantic tag-based profile has the following features:

- *Tag weight* represents the tag importance for this user. The most common way to calculate the tag weight is to use tag frequency. The more frequent a tag has been used, the greater the tag weight.
- *Link weight* represents the relationship importance between two tags for this user. In our thesis, we propose the semantic-based co-occurrence and the social wisdom to enhance the pure co-occurrence method.
- *Tree views profiles* represent the profile from different aspects. Using self tags we can show the most subjective opinions about this user; using tags of all users

we can show the most objective opinions about this user; using friends' tags (or the tags of a group of users) we can show this community's opinions about the target user. [8] Thus, we identify three viewpoints: personal, social and global to show the different aspects of a person.

5.2 Our Idea

Tag clouds represent a set of tags as weighted lists. The more often a tag has been used, the larger it will be presented. This mechanism can be used for tag profiles, through which people can quickly skim through the characteristics of a user. We use font size and color to emphasize the weights of tags. Unlike traditional tag clouds which are 1-D lists, we use a force-directed layout to present weighted tags and their links in an aesthetically pleasing way. Forces are assigned on the set of nodes (tags) and the set of edges (links). The whole graph is then simulated as a spring system which quickly comes to a stable state. The layout is shown in Figure 5.1.

In order to display the structure of the tag-based profile, we use radial layout to represent the semantic relationship between one target tag and its relevant tags. When a tag is clicked on in this graph, it becomes the target tag and is placed in the center of a series of concentric circles which are composed of the target tag's relevant tags. In this layout, two degrees of separation from the target tag are shown. (See Figure 5.2)

An iTunes-styled coverflow design is incorporated with a 3D carousel effect (See 5.3). Users are able to switch amongst different views of the tagging-based profile in an easy and quick fashion.

(a) Tag profile with cloud status (b) Tag profile with graph status

Figure 5.1: Force-directed layout on tag profile visualization.

(a) ``tagging" is the center of tag profile (b) ``del.icio.us" is the center of tag profile

Figure 5.2: Radial layout on tag profile visualization.

Figure 5.3: iTunes-styled 3D Carousel Coverflow

5.3 Profile Presentation From Three Viewpoints

Based on a user's content collection, we can utilize self tags to construct his/her user profile. These tags represent a user's interests and knowledge subjectively and we can call such a profile the ``personal view of a user profile.'' On the other hand, the tags of a group of people can represent the opinions of this community on the same content collection. In this thesis, we define the members of this group as the user's friends. Friends' tags can form the ``social view of a user profile.'' Finally, we extend the members of a group to all users in the system. The tags of all users in the system can represent general, objective opinions on the same content collection and form the ``global view of a user profile.''

The tag-based profile presentation from the personal view of a sample user is shown in Figure 5.4. The user's interests can be identified quickly by looking at this user profile. With much ease we can easily identify that this user may like documents about ``visualization,'' ``design,'' and ``flash.'' Consolidating this information, we can probably infer that this user may be a researcher whose research topics may be about ``tagging'' and ``visualization.'' We may also infer that she is interested in learning technologies

like ``flash" and ``flex."

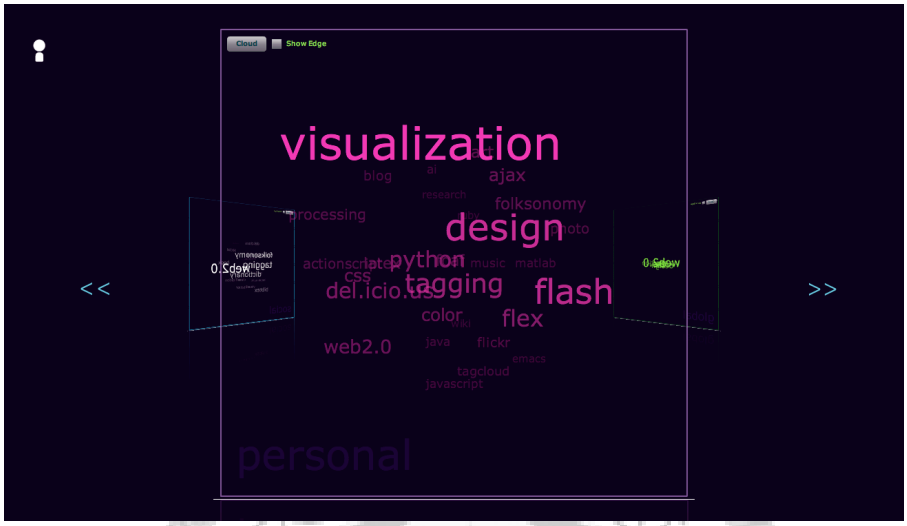


Figure 5.4: Tag-based profile from personal view

Having observed the user profile from his/her personal view, we can easily switch to see the same user profile from the social view and the global view, which are respectively illustrated in Figure 5.5 and Figure 5.6. Through these two views, we can learn more about the user by considering the opinions of the user's friends or the general public. One interesting thing is that tags from the social view and the global view are more general than those from the personal view. We can identify broader knowledge and opinions on these contents.

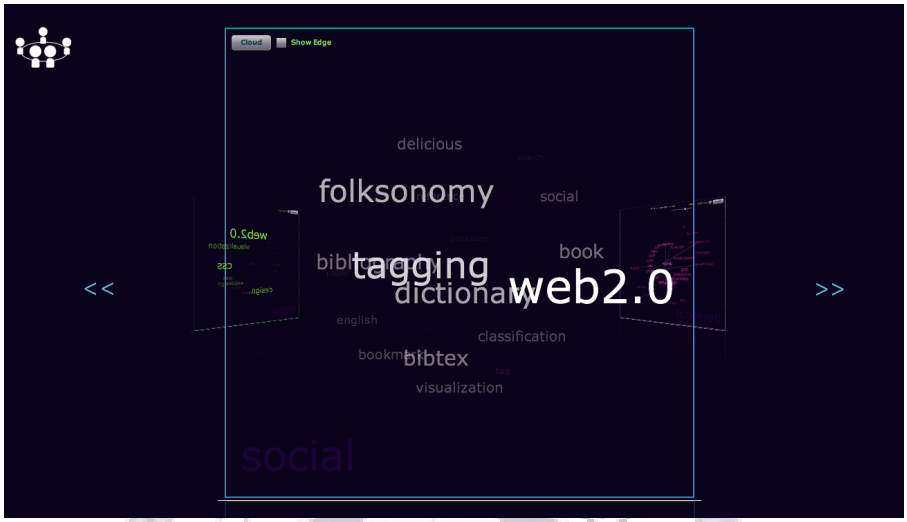


Figure 5.5: Tag-based profile from social view

Figure 5.6: Tag-based profile from global view

Chapter 6

Experiment and Evaluation

6.1 Data Collection

In our work, we use social media data to profile a person. Therefore, we crawl data from del.icio.us[1], a popular social bookmarking website. A user's data on this site is often in the form as illustrated in Fig 6.1. A bookmarked item includes URL, title, timestamp, popularity (which means the number of people who have also bookmarked the item), and tags (which are given by the user). In addition, we crawl the tagging history on a specific URL. (See 6.2)

We utilize tag analysis to obtain weighted tags in three viewpoints [8] and calculate the semantic relationship between tags based on the approach described in Chapter 4. We design a visualization (5) to present the resulting semantic tag-based user profile.

[1] http://del.icio.us

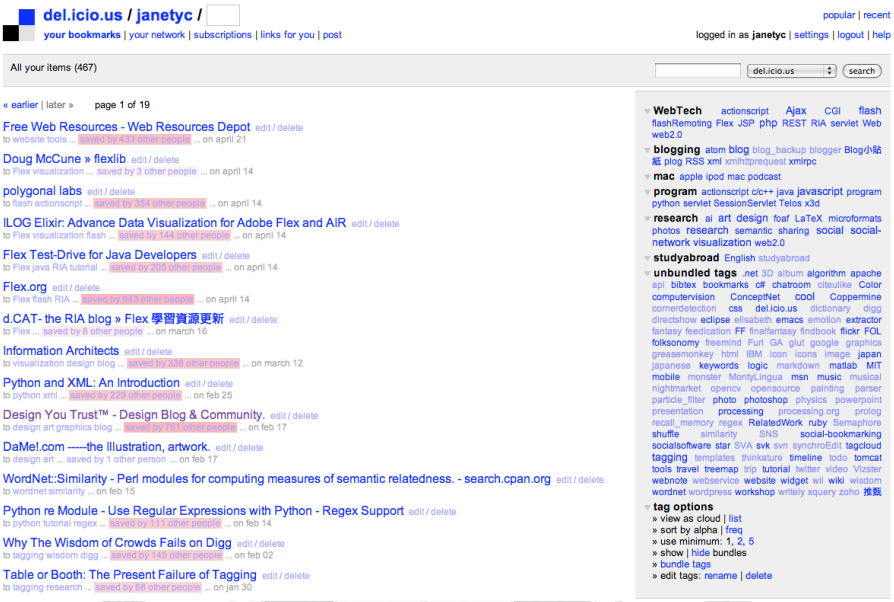


Figure 6.1: A person's bookmark collection on del.icio.us.

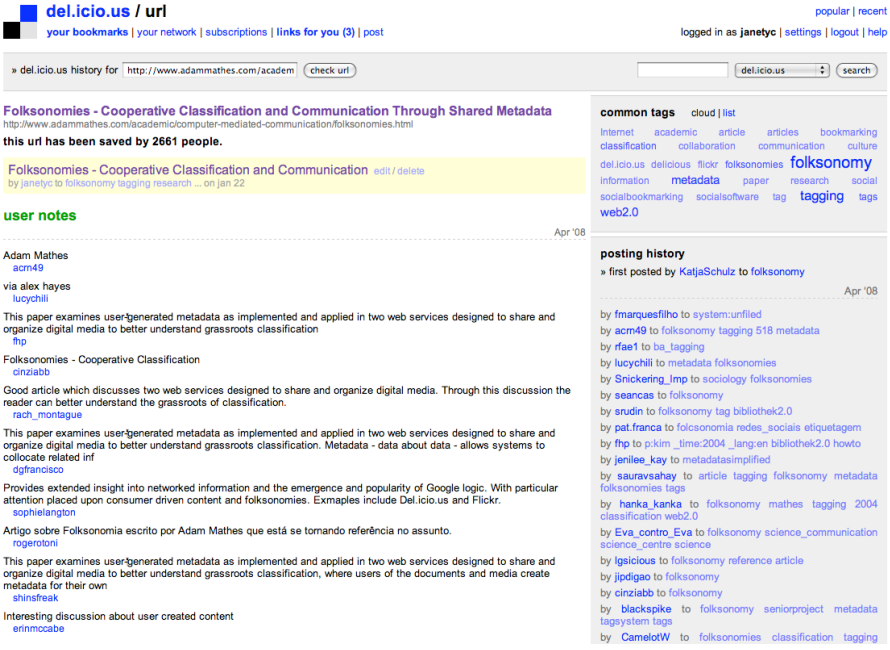


Figure 6.2: All tags and comments for a specific URL.

6.2 User Study Design

We design a user study to evaluate if the semantic relationships indeed fit a user's knowledge and preferences. The study starts with a simple task for the user, where he/she, given a randomly picked target tag and a list of his/her top 50 most used tags, is asked to choose from the list at least 3 other tags that are considered relevant (based on the user's knowledge and preferences) to the given target tag. The same task is repeated five times, each time with a different randomly picked target tag. The screenshot of Experiment is shown in Figure 6.3.

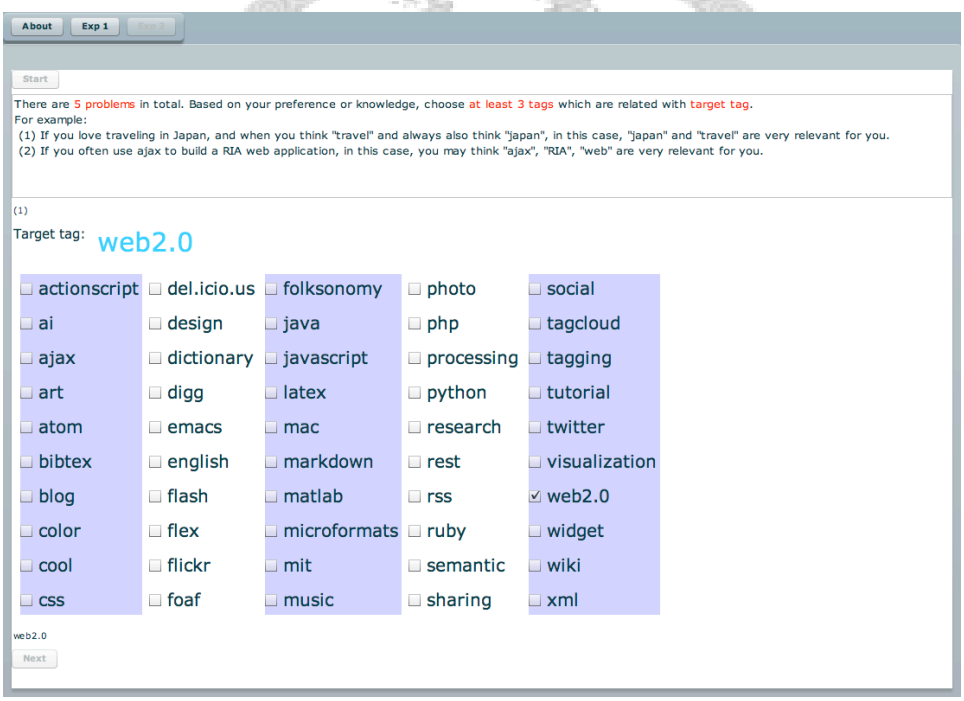


Figure 6.3: Screenshot of Experiment.

A small-scale user study is conducted with our lab members and strangers, in which data from ten different individuals is collected. There are 50 related tag sets and 242 distinct tags in total. These tags as ground truth represent the ten individuals' true

knowledge and preferences. Once the data is collected, we plan to compare the following five methods:

- Co-occurrence: Only use the base co-ccurrence method to construct the personal knowledge network.
- Co-occurrence + Social Wisdom: After using the social wisdom to reinforce the links of tags and URL, we construct the personal knowledge network by co-occurrence.
- Semantic-based Co-occurrence + Social Wisdom: After reinforcing the data, we utilize the concept-based co-occurrence method we proposed to construct the personal knowledge network.
- WordNet: Only use WordNet to construct the strength of the semantic relationships between tags and to create the WordNet-based knowledge network.
- ConceptNet: Only use ConceptNet to create the ConceptNet-based knowledge network.

Given the five knowledge networks for each individual, we use spreading activation to obtain 5 related tag sets for a target tag. We calculate the precision and recall of the 5 methods respectively. The precision and recall are defined in Equation 6.1. We define $RelatedTags$ as a set of related tags produced by our methods and $AllRelatedTags$ as a set of related tags collected by the user, as well as our ground truth. We use balanced F-score to combine precision and recall in Equation 6.2.

$$\begin{aligned} Precision &= \frac{|RelatedTags \bigcap AllRelatedTags|}{|RelatedTags|} \\ Recall &= \frac{|RelatedTags \bigcap AllRelatedTags|}{|AllRelatedTags|} \end{aligned} \tag{6.1}$$

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6.2}$$

6.3 Experiment Results

Our result is shown in Figure 6.4. We can see that co-occurrence methods (the first three methods) perform better than the WordNet and the ConceptNet methods. This implies that co-occurrence better captures personalized preferences and retrieves broader knowledge from the human brain. On the other hand, WordNet and ConceptNet are limited to capturing the meaning of lexicon and the common sense knowledge.

To compare the co-occurrence methods, we discover that the co-occurrence with social wisdom method gains the highest value in recall. This result is reasonable and conforms with our expectation. Social wisdom can reinforce missing links between tags and URL. We do not enrich with tags that the user has never used because we want to model the user's knowledge with his/her own vocabulary. Reinforcement makes the relationship between tags stronger. However, it also inevitably introduces some noise. Compared with the co-occurrence with semantic and social wisdom method, a decrease of precision indicates some noise is present. In terms of precision, we discover that the highest value belongs to the semantic-based co-occurrence with social wisdom method. Semantic similarity helps to filter out some noise which is produced by the social wisdom.

We use the F-score to calculate the weighted harmonic mean of precision and recall for the different methods. The result of the F-score is shown in Figure 6.5. As a whole, the co-occurrense with social wisdom method and the semantic-based co-occurrence with social wisdom method perform better than others. Particularly, the WordNet and the ConceptNet methods perform most poorly.

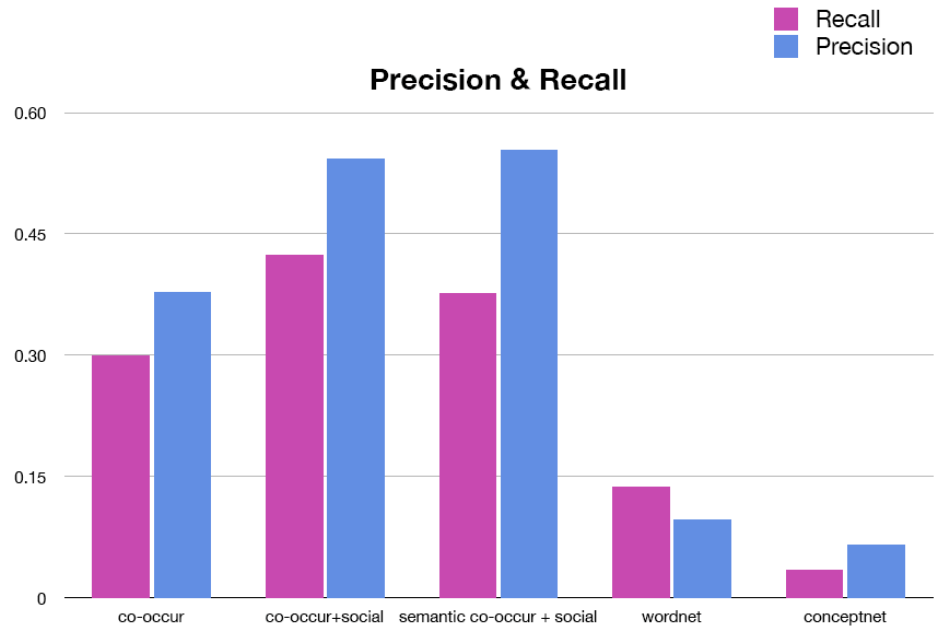


Figure 6.4: Precision and Recall on experiment.

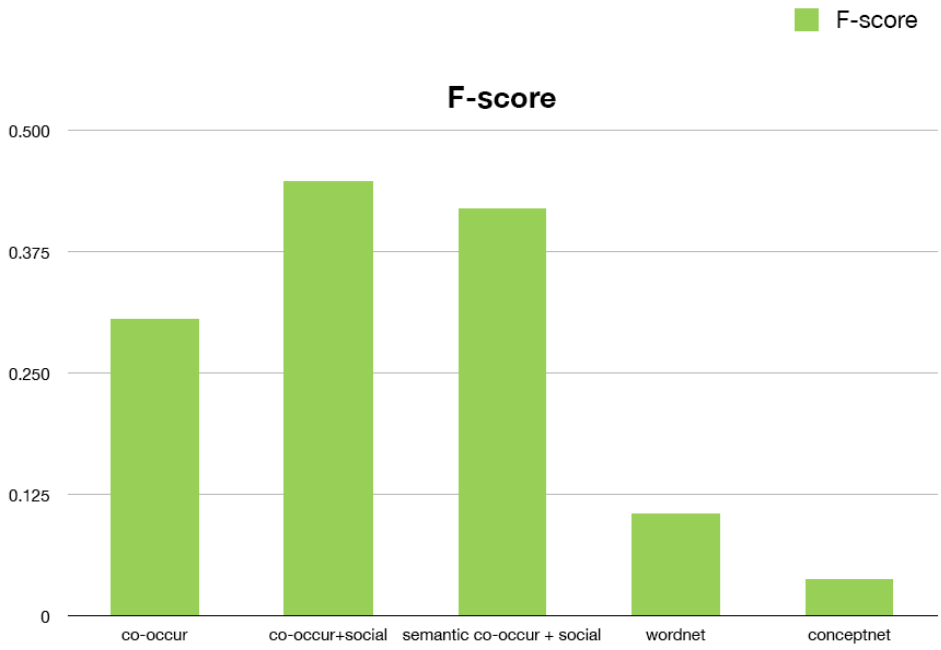


Figure 6.5: F-score on lab members

As we previously clamed, social wisdom is useful for lazy users. Figure 6.6 illustrates this argument for one particular test user. The two pictures in the figure use two target tags: travelagent and bibtex. In the picture on the left, ``travel,'' ``hotel,'' and ``taiwan'' are not associated with the target tag ``travelagent'' when the co-occurrence measure is used. However, when social wisdom is applied, these tags become associated with the target tag. In reality, this user truly collects webpages about ``travel,'' ``hotel,'' ``taiwan,'' and ``travelagent.'' Another example with the target tag ``bibtex'' is shown in the picture on the right. We can see that ``latex,'' ``thesis'' and ``writing'' are associated with ``bibtex'' when social wisdom is applied.

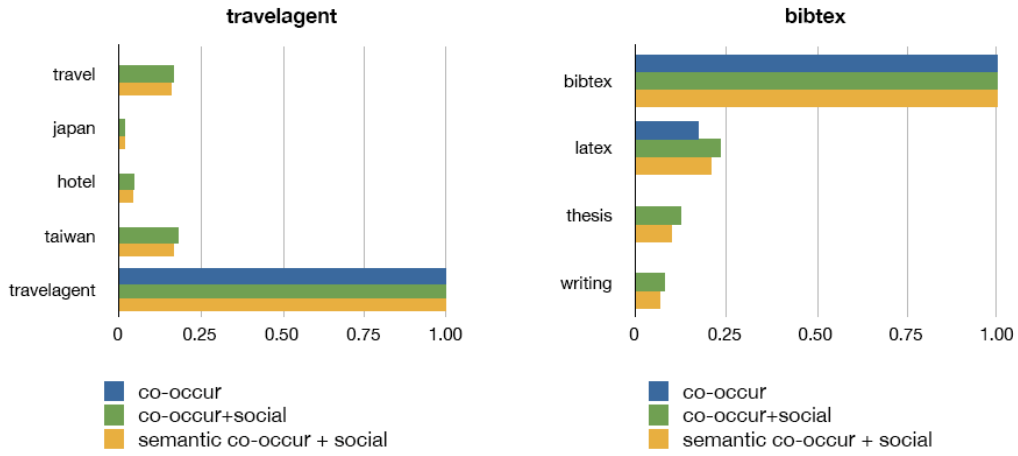


Figure 6.6: Semantic relationship of ``tavelagent'' and ``bibtex.''

Chapter 7

Conclusion

In this thesis, we propose a novel approach to user profiling based on the tags associated with one's social media. Any user can be profiled as a set of weighted tags with semantic relationships of different strength. Weighted tags can represent the user's interested topics and preferences, while tag relationships can represent the relationships between the topics.

This thesis defines three types of knowledge and three retrieved approaches on analyzing social media data. Co-occurrence can reveal personalized association and social wisdom can reinforce the tagging data. Common sense computing is used to construct the global knowledge for calculating semantic similarity between tags. Furthermore, we integrate the idea of common sense computing and the co-occurrence measurement and propose the semantic-based co-occurrence method to calculate the strength of the semantic relationship in a tag-based profile. Based on the relationship between tags, we can model a user's knowledge and preference and capture the user interests in more detail. In addition, we design a presentation tool with a 3D carousel effect to show the tag-based profile from three viewpoints. In the end, we design a user study to collect the real opinions for evaluating whether the semantic similarity actually fits the user

preferences.

7.1 Summary of Contributions

This thesis proposes a novel approach to constructing an individual's personal profile using his/her collection of media contents automatically and effectively. We propose the semantic tag-based profile, which is profiling an individual as a set of weighted tags with semantic relationships. Furthermore, we propose a profile presentation which visualizes the user's interested topics, as well as the preference degrees and semantic relationships between topics. The contributions of this thesis are summarized as follows:

- Semantic tag-based profile can model an individual's knowledge and interests.
- Definitions for three types of knowledge and their corresponding retrieved approaches using a user's bookmarking data. Co-occurrence reveals personalized association. Social wisdom utilizes community knowledge to reinforce incomplete data. Common sense computing builds global knowledge on these tags from daily life or dictionary.
- *Semanitc-based co-occurrence* measurement integrates the advantages of common sense computing and co-occurrence. It is used to compute the strength of the semantic relationships between tags and to capture the relationships between a user's interested topics.
- Experiments involving five people to collect real opinions and knowledge of users. The result shows that our methods better fit the users' knowledge and thoughts.

- A profile presentation with a 3D switch effect is designed to visualize the semantic tag-based profile from alternative viewpoints.

7.2 Future Work

User profiling is the basis for many applications especially for recommender system. Based on a user's profile, the system could recommend items (books, webpages, products, etc.) of interest to the users. Using the semantic tag-based profile, recommender system can provide more relevant products or services to satisfy the user needs.

The semantic tag-based profile may be useful for other applications. Another interesting application is for job search or matchmaking . We can provide a service to automatically help profile viewers find the the right person. For job search, job seekers can provide their profiles and recruiters can browse through these profiles to identify qualified candidates for the jobs. For matchmaking, depending on the purpose of a search, different match criteria should be considered. The matchmaker may define different *match functions*, such as *similar*, *complementary*, or *having interests overlap*, which should be further explored in our future work.

Bibliography

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971--980, New York, NY, USA, 2007. ACM Press.

[2] E. Bertino, G. P. Zarri, and B. Catania. *Intelligent Database Systems*. Addison-Wesley Professional, 2001.

[3] A. Collins and E. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407--428, 1975.

[4] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, March 1998.

[5] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems, Aug 2005.

[6] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198--208, April 2006.

[7] Y. Hasan-Montero and V. Herrero-Solana. Improving tag-clouds as a visual information retrieval interfaces. In *Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies*, October 2006.

[8] Y.-C. Huang, C.-C. Hung, and J. Y.-j. Hsu. You are what you tag. In *Proceedings of AAAI 2008 Spring Symposium Series on Social Information Processing*, Stanford University, California, March 2008.

[9] B. Kerr. Tagorbitals: a tag index visualization. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Sketches*, New York, NY, USA, 2006. ACM.

[10] H. Liu and P. Maes. InterestMap: Harvesting social network profiles for recommendations. In *Proceedings of the Beyond Personalization 2005 Workshop*, 2005.

[11] H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22, 2004.

[12] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31--40, New York, NY, USA, 2006. ACM.

[13] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004.

[14] E. Michlmayr and S. Cayzer. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)*, May 2007.

[15] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database*. *in International Journal of Lexicography*, 3(4): 235--244, January 1990.

[16] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288--299, June 2007.

[17] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*, pages 1024--1025, San Jose, CA, July 25-29 2004.

[18] R. Sinha. A social analysis of tagging. World Wide Web electronic publication, 2006.

[19] M. Stefaner. Visual tools for the socio--semantic web. Master's thesis, University of Applied Sciences Potsdam, June 2007.

[20] J. Voss. Tagging, folksonomy & co-renaissance of manual indexing? *Arxiv preprint cs/0701072*, 2007.