A PERSONALIZED RECOMMENDER AGENT FOR THE WORLD WIDE WEB – A SEMANTIC PERSPECTIVE

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

DECEMBER 2010

By

Pei-Chia Chang

Dissertation Committee:
Luz M. Quiroga, Chairperson
Kim Binsted
David N. Chin
Daniel D. Suthers
John Zuern

Keywords: personalization, ontology, Wikipedia

To My Father

&

Heavenly Father

Acknowledgements

It is a pleasure to thank those who made this dissertation possible. My advisor, Dr. Luz Quiroga, contributed a significant portion of her time to guide me in research methods, conference publication and dissertation writing. Without her, I would not have been able to reach this point.

I am also grateful for the suggestions from my committee members. Dr. David Chin has always been helpful in providing me constructive feedback, travel grant support, and conference information. As my Master program's advisor, Dr. Dan Suthers has seen my learning from a naive student to a young researcher. I appreciate immensely his contribution to the process. Additionally, Dr. Kim Binsted has been critical to identify the position of my work across domains. Last but not least, I owe great gratitude to Dr. John Zurn, who spent his time in understanding and commenting this work in a field that is different from his past concentration.

Furthermore, many thanks must be given to this study’s participants who volunteered out of courtesy. I was so blessed to receive such kind help from them. Certain credits should be given to Dr. Mei-Chih Lai and Yanhan Nie who helped me enormously for this dissertation. Lastly, I would like to thank my parents, my family and my institute (Writing Center, Graduation Student Organization, Center on the Family, and University of Hawaii Foundation) for their support.

Abstract

Web personalization aims to provide useful Internet information and avoid information overload. Most web personalization studies focus on a single website. Recommending pages across websites is challenging due to a variety of concerns, such as dynamic web and diverse interests, conflict of interest, shared communication protocols required, and model reusability.

This work explores the potential of augmenting Wikipedia's categories with page keywords as an agent system for semantic user modeling to recommend pages across websites. The recommender agent focuses on modeling individual web users' topical interests, using the content-based usage analysis at the client-side. The system also promotes serendipity ─ novel and interesting information ─ as a major factor in our recommendations by considering the coverage of a user's interests via the Diversity Index using the categorical topology.

This dissertation's main research question is "Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?" Three sub-questions were investigated sequentially in the computer science domain. The investigation examined the system's core components separately, and tuned up individual components before investigating the overall performance of recommendations.

1. Can our content model correctly identify the topics of a web page?
2. Does our content-based user model semantically capture a user's interests?

3. Does Diversity Index measure the coverage of a user's interests in the computer science domain?

This study evaluated the system's performance on recommendations regarding topicality and serendipity with 25 professionals as participants in the computer science domain. Results indicate that our system's performance is slightly better than the pure content-based vector space model (VSM) regarding topicality, and significantly better regarding serendipity. A further investigation reveals that our system is able to identify serendipitous recommendations that VSM may fail to recommend. The system's superior performance in serendipity is possibly due to the augmentation of Wikipedia's categories with keywords as well as the utilization of the categories' topology.

This work is significant for four reasons. First, it emphasizes the convergence between content modeling and user modeling by means of augmenting Wikipedia's content and usage mining. Second, using the semantics (vocabulary, categorical association, etc.) of Wikipedia for user modeling considering serendipity is worthwhile as the factor is not addressed extensively in the literature. The model is deliberately constructed as a research platform based on heuristic information extraction on keywords and allows for more heuristics. Third, a user's topical interests are modeled using Wikipedia's categories, which yield a simple model that can be interoperated among different websites. The model, with its simplicity, is at the client side, allowing more user control and reducing privacy concerns. Fourth, a methodology is supplied to researchers for further development of similar recommender agents.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1. Motivation

Topic Introduction

Few studies have addressed cross-website page recommendations, which is this dissertation's focus within the field of web personalization (WP) research. This dissertation emphasizes that eliciting user interests among different topics within a domain is an important concern in cross-website page recommendations. By enhancing Wikipedia's categorization system with heuristic information extraction, this work constructed a recommender agent system that promotes serendipity in domain specific recommendations. Serendipity is defined as providing novel and interesting information. It is a new dimension that caught researchers' attention. The system also assumes topicality in its recommendations. Topicality means providing topically relevant information to a user's interest. The system was expected to provide serendipitous recommendations that are topically relevant to a web user's interests according to the usage data. Thus, we compared the system's performance regarding topicality and serendipity with the classical vector space model in the computer science (CS) domain. Results indicate a superior performance of the system, especially regarding serendipity.

Eirinaki and Vazirgiannis, define web personalization as "the process of customizing a web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the web context, namely, structure, content, and user profile data" (Eirinaki & Vazirgiannis, 2003, p. 1). Cross-system personalization (CSP) aims to provide personalization based on profiles and protocols among different service systems (Mehta, Niederée, & Stewart, 2005). CSP for websites is our focus. A website is defined as pages belonging to the same domain name based on URL.

Motives

Although WP aims at improving the user experience from a website provider's perspective, we intend to focus more on the client user's perspective and address serendipity for personalization. For example, we intend to present users a thematic view of pages from different websites according to user preferences rather than generating page elements dynamically for a single website. Manual customization is a conventional way to achieve personalization but it is tedious. We propose that Wikipedia – the largest and dynamic knowledge base – and available machine learning techniques offer better ways to complement the information needs of cross-website personalization.

Vision/Scenario

In view of the above motivation, the vision of the proposed agent system is described as follows. Imagine that an agent-like system retrieves information from different websites and filters websites according to a user's domain interests. The system continually evolves through web usage and content mining based on the user's web browsing behavior. It not only helps the user locate a previously accessed document, but also acts as a personalized web crawler, seeking new information that may be of interest to the user. Our ultimate goal is to recommend informative web pages that satisfy the user's needs and promote serendipity in the recommendations with minimal user effort.

The following two paragraphs describe two scenarios in the CS domain– Leo, a programmer and Stephanie, a research assistant – that can be fulfilled of the proposed agent system. The case of Leo demonstrates the usefulness for the agent system to identify interesting pages related to constantly updated topics that he is not aware of. Stephanie's scenario highlights the possibility for the agent system to identify the overlapping topics of her interest. Both scenarios illustrate the focus of serendipity in addition to topicality in recommendations.

Leo is a programmer who works on mobile programming. He is interested in technical gadgets, portable devices, and mobile applications. To keep himself updated, Leo installs the recommender agent system in his computer. Periodically, the agent recommends him web pages that introduce new technical gadgets or review platform-specific applications. Even though Leo regularly visits his favorite websites, he finds the agent is helpful in identifying newly created pages or websites that he is not aware of, and that pertain to his interests.

Stephanie is a research assistant who specializes in telecommunication. Her job requires the investigation of various wireless technologies. She needs to collect magazine articles, conference publications and journal papers in order to write reviews and analyses for the different research projects she participates in. To fulfill her job's needs, Stephanie subscribes to mailing lists and uses Google scholar for looking into information. She installs the recommender agent system at her friend's suggestion. Stephanie is happy about the pages recommended by the agent. From time to time, certain recommended pages are related to a combination of different research projects Stephanie is involved in. To Stephanie's surprise, the agent system is able to identify the overlapping parts of her research topics and recommend her informative pages quite often.

Statement of purpose

The purpose of this dissertation is to develop and test a dynamic approach to capture a user's topical interests for CWP mainly through content-based user modeling. We adapt Wikipedia's content (categories and keywords) to derive an ontological model and apply usage mining based on the model to bridge the gap between content modeling and user modeling. Additionally, the topology of Wikipeida's categories is used to identify the coverage of a user's topical interest within a selected domain. This is intended to promote the topical diversity in CWP. We aim for a dynamic model due to the highly changeable page content and various user interests in the web environment.

Theoretically, we intend to propose an approach that is fine tuned for the web environment according to rationales and constructs in information retrieval and user modeling, such as information foraging theory (Pirolli & Card, 1995) and ontology (Gruber, 1995). Moreover, we foresee the implication of different service domains (e.g. expertise recommendations) or knowledge domains (e.g. medical informatics) of utilizing such a dynamic model because it is highly adaptable to a constantly changing web environment.

1.2. Statement of Problems

Time-Consuming Searches

WP potentially saves both users' and experts' time. A website may contain all types of knowledge, but only a few are of interest to a user. Information foraging theory (Pirolli, 1997, 1998, 1999; Pirolli & Card, 1995) explains the process by which a user explores and identifies desirable information. It is based on certain principles and literature that Pirolli refers to: ecological food-foraging strategies, information scent (Huberman, Pirolli, Pitknow, & Lukose, 1998), maximizing valuable information and optimal-foraging theory (Stephens & Krebs, 1986). Basically, one makes decisions based on matching one's information goal with the text on a display screen. Wrong decisions result in frustrating and time-consuming backtracking. Such problems arise because *users may not know exactly what they are looking for.*

The Anomalous State of Knowledge (ASK) (Belkin, 1977, , 1978; Belkin, Brooks, & Oddy, 1979) supports this point. It describes information seekers' cognitive state of knowledge structure. Relating to ASK, Bate (Bates, 1989) uses an analogy of "berrypicking" to describe the browsing and refining process, which may take some time. In addition to user's time, WP often makes use of an expert-constructed knowledge base, which is manually built and time-consuming as well.

Incommensurate Mental Models

Differences among mental models (Gentner & Stevens, 1983; Johnson-Laird, 1983; Norman, 1983; Payne, 1991) offer one possible explanation of the mismatch between user prediction and screen texts. The language novices use may be different from what experts adopt. In their study, Furnas, Landauer, Gomez, and Dumais (1987) have shown the choice of vocabulary is a problem in human-system communication. Similarly, what a user formulates as keywords reflecting his thoughts may turn out to differ from those that are generally accepted ones.

Dynamic Web and Diverse Interests

Another problem is the dynamic nature of the web environment, meaning that information changes constantly. Not only does information change, so do a user's interests and information needs. Furthermore, with the popularity of social-networking websites, users not only find information, but also discover interesting people from the World Wide Web (WWW). Therefore, in the digital age of information overload, a filtering mechanism for individuals is certainly necessary. A timely recommendation of updated information could potentially align the user's interests with the nature of dynamic web. Additionally, other factors, such as novelty or serendipity of recommendations, deserve considerations to better satisfy various user interests or expectations (Herlocker, Konstan, Terveen, & Riedl, 2004).

Serendipity is a primary focus of this dissertation but its evaluation may not be straightforward. Herlocker et al. comment that "designing metrics to measure serendipity is difficult, because serendipity is a measure of the degree to which the recommendations are presenting items that are both attractive to users and surprising to them."(Herlocker, Konstan, Terveen, & Riedl, 2004, p. 43) Shani and Gunawardana suggest asking users to mark unexpected recommendations and examine if they follow the unexpected recommendations, as the behavior will be interpreted as reviewing interesting or successful recommendations (Shani & Gunawardana, 2009). Nevertheless, this approach assumes that following recommendations implies a success. Therefore, there seems to be no golden standard for measuring serendipity yet. As a result, this dissertation borrows the definitions of novelty and interestingness, which are defined in section 4.2.

Privacy Concerns

Privacy is a priority issue raised by users. As summarized by Kobsa (2007), Internet users are reluctant to disclose data about their Internet behavior, transactions and detailed demographic information. Even though there is an increasing awareness of the need to enhance the privacy protection of Internet users, suggestions and regulations still cannot ensure users' confidence.

Lack of Effective Cross-Website Personalization Systems

Most WP studies, as stated in the definition, focus on customizing a single website. and few studies have addressed cross-website personalization. Nevertheless, by modeling usage data from multiple websites, user behavior can be predicted with a higher accuracy (Padmanabhan, Zheng, & Kimbrough, 2001). Padmanabhan et al. refer to this approach as "user-centric," as opposed to "site-centric," which only makes use of data from a single website. In this dissertation, personalization based on user-centric data from different websites is named cross-website personalization (CWP) to better fit our research scope. Cross-system personalization (CSP) is the common term used in literature.

CSP poses three challenges: 1) modeling users in extensible and unified ways that can be interpreted and used by different systems; 2) collecting and filing the data of user interactions and updating user profiles based on the unified user context model; 3) developing the conceptual and technical methodology for different systems to make use of the collected user profiles and for improved support of the users (Niederée, Stewart, Mehta, & Hemmje, 2004). In addition to these difficulties of model integration, conflict of interest from competitive personalization systems is another challenge.

Similarly, Zhang, Song and Zhang identify two deficiencies of existing approaches (2006). One is the model replacement from one user-adaptive system to another. The other is the lack of agreement on a shared communication protocol among agents or personalization systems. Zhang et al.(2006) further provide an overview of four different model reusing approaches in service architecture on a web basis. These approaches include 1) generic user modeling system, 2) multi-agent, 3) ontology and 4) unified user context. Generic user modeling systems separate personalization engines from service applications (Kobsa, 2001). In a multi-agent platform, a model can be either universally accessed or dispersed as fragments. Ontology defines the interpretation or specification of a shared standard. Having a unified user context requires a mediator (Frommholz, Mehta, Niederée, Risse, & Thiel, 2004) or a common protocol (Heckmann & Krueger, 2003) in the process.

Cross-system personalization is in an early stage, but is emerging due to the popularity of distributed systems (Dolog & Nejd, 2003). Furthermore, CSP may need to concern different aspects of a similar topic or even across topics, because the focus of various comparable systems may differ. This variation in focus raises issues beyond content similarity, probably including topical diversity as well.

The problems identified above motivate us to envision the possibility of an adaptive user model that reflects a user more extensively and can be reused by different servers. Current model reusing approaches provide us certain research directions. The next section presents background knowledge required for the discussion of our research question and scope ─ agent, web personalization and ontology.

1.3. Background & Definitions

Agent

"An agent is an autonomous software entity that – functioning continuously – carries out a set of goal-oriented tasks on behalf of another entity, either human or software system. This software entity is able to perceive its environment through sensors and act upon it through effectors, and in doing so, employ some knowledge or representation of the user's preference (Wooldridge, 1999)." (Andreas & Pericles, 2005, p. 42) Agents need to be trained in order to react with the above functionalities. Existing systems commonly adopt WP procedures, which are explained next.

Web Personalization: Processes, Data, System Techniques, and Our Focus

According to Mobasher, Cooley, and Srivastava, a typical WP process includes: 1) the modeling of web objects (resources that can be retrieved via websites, such as pages) and subjects (users); 2) the analysis of web objects and subjects; and 3) the matching and recommendation of web objects and subjects (2000). WP is therefore relevant to the research communities of information retrieval, user modeling, web mining, and recommendation systems.

Recommendation systems (Budzik, Bradshaw, Fu, & Hammond, 2002; Kautz, Selman, & Shah, 1997; Konstan et al., 1997; Mooney & Roy, 2000; Reichling, Schubert, & Wulf, 2005; Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994; Resnick & Varian, 1997) are common applications of WP. Information retrieval, user modeling, and web mining are closely-related research areas that study the correlation between web data and web users. According to Srivastava, Cooley, Deshpande, and Tan (2000), there are four main categories of web data: content, structure, usage, and user profile. Web users could be an individual or a community group.

As far as the underlying techniques are concerned, WP systems fall into four major categories: 1) manual decision-rule systems, 2) collaborative-filtering systems, 3) content-based-filtering agents, and 4) usage-based systems (Eirinaki & Vazirgiannis, 2003). In addition, a recommended item could be either web data (Konstan et al., 1997; Mooney & Roy, 2000; Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994) or web users (Budzik, Bradshaw, Fu, & Hammond, 2002; Kautz, Selman, & Shah, 1997; Reichling, Schubert, & Wulf, 2005).

Regarding our focus, the combination of **web content and usage mining** is a prominent WP research direction, which is discussed in the literature review (chapter 2). Recent studies focus a lot on collaborative filtering and web usage mining. However, the content-based approach is an essential foundation for studying user behavior to gain better semantic interpretations. Moreover, content analysis complements usage mining, especially for newly-added web content. We particularly address the use of ontologies as our selected approach for content mining.

Ontology

“An ontology is an explicit specification of a conceptualization.” (Gruber, 1995, p. 908) Regarding conceptualization, Gruber defines “A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose.” (Gruber, 1995, p. 908). For example, an “is-a” concept can be defined in the specification to represent the hierarchal property of a taxonomy. Similarly, contextual or structural associations can be formalized as well. Formalizing concepts in a relationship to one another may be used for knowledge inference in general and specific situations. The use of ontologies is thus a key technique that characterizes the semantics of information exchange (Fensel, 2003, p. 4).

1.4. Scope and Research Questions

Research Path

The major effort of this dissertation focuses on an analysis of the relations between web pages and the individual user through Wikipedia’s content in the CS domain. The domain was selected because that there are relatively rich data in CS compared to other domains. For the purposes of minimizing confounding variables (e.g. topical diversity) that may affect the system performance, only a single domain was selected. Additionally, while there are many formats of web pages, we intend to model knowledge only for textual web pages. This work specifically addresses the typical paradigm for deriving a user model for text filtering, the source of model derivation, the duration of a model, and certain attributes of the model to deal with the previously-mentioned problems.

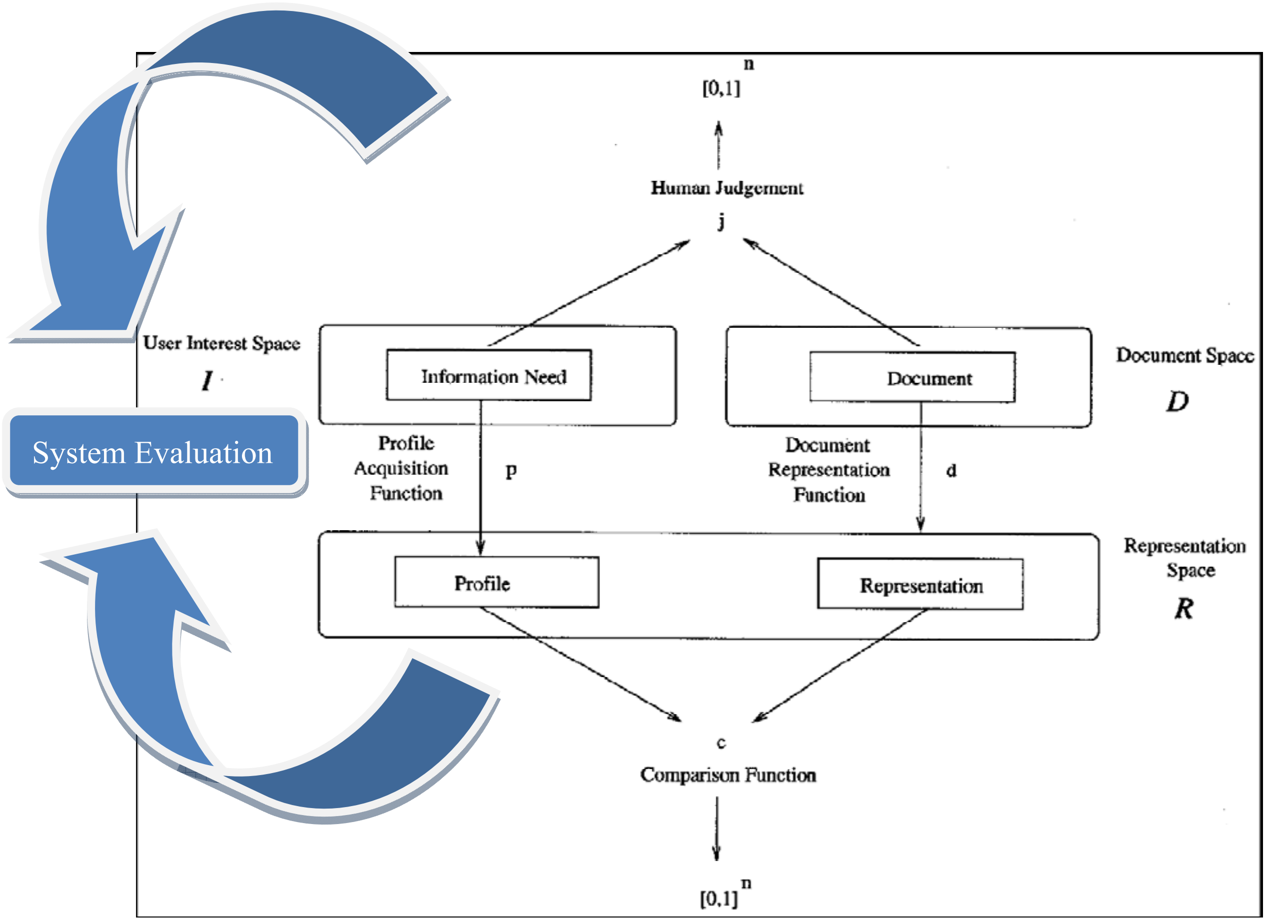

Figure 1.4-1 Text Filtering System Model from D. W. Oard with Adaption

A typical process for text filtering systems perfectly describes our research path. Figure 1.4-1 is the model for text filtering systems that Oard proposed (1997). We added "System Evaluation" on the left. User interest space *I* and Document space *D* correspond to user modeling and content modeling respectively. We intend to program the profile acquisition function *p* and the document representation function *d*. Internally, an automated process performs the comparison function *c* to find the best matches between user interests and page content in our agent system. Eventually, the comparison function is evaluated using human judgment, which is emphasized as system evaluation in the left part of the graph.

Individual Client-side Modeling in the Long Term

Specifically, we attempt to recognize a user's web access behavior through modeling usage sources that are readily available at the *individual client (user) side*. The client side provides rich information to understand a user's preferences in the *long term*. In addition, a user may be hesitant about giving out private usage information online, and may be more inclined to disclose personal data at the client side (Kobsa, 2007).

Model Reusing or Compatibility

Our focus has been identified, but model compatibility is one more concern. WP means something different to everyone, which implies various means of personalization services. Ample modeling services are available, such as email recommendations or personalized web portals. In view of these existing options, instead of creating a user model from scratch, model reusing or unifying has become an alternative way to build a WP service across different platforms (Gonzalez, Angulo, Lopez, & Rosa, 2005; Heckmann, Schwartz, Brandherm, & Kröner, 2005; Kobsa, 2001; Niederée, Stewart, Mehta, & Hemmje, 2004; Zhang, Song, & Zhang, 2006). Therefore, we need to keep in mind designing a system that facilitates model compatibility and extensibility.

Semantic Modeling

Reusing models can also provide a semantic view of a modeled user based on the assumption that a user's behavior is consistent when browsing domains of his interest. For example, the profile of a user's reading preference from Amazon.com can be combined with the user's bookmarks and tags in Delicious[1] for reuse. The reason is as follows: viewing the service content of the two websites from a semantic perspective, it may be possible to find that the user has interests in certain knowledge domains. For the user, a book and a bookmarked web page are just different representations of the user's topical interests. Therefore, from the example, what seems promising is that *WP should focus more on the semantic level regardless of the service media.*

Research Foci and Hypotheses

Our research addresses each point of the above concerns. We started with three foci that guide our literature review, which are as follows:

F1. Content modeling for web pages, *addressing the semantic perspective*

F2. Modeling a web user's topical interests, *reducing user effort, controlling privacy and concerning diversity*

F3. Matching page models with a user model, *promoting serendipity*

Our main hypothesis states that user models shall align with content models regarding semantic relevance. This statement implies that the content model shall carry semantic representation, as is explored using Wikipedia's content in our first focus. We verify our main hypothesis by coherently combining user-provided and content-derived information to make a semantic user model. Other hypotheses are stated in section 4.3. Our second focus examines this aspect by modeling a user's topical interests from the usage data at the client side to grant users more control. In addition, we consider the

[1] http://www.delicious.com

topical diversity of the usage data. A last hypothesis is that applying implicit modeling approaches based on Wikipedia's content yields serendipitous recommendations. "A serendipitous recommendation helps the user find a surprisingly interesting item he might not have otherwise discovered." (Herlocker, Konstan, Terveen, & Riedl, 2004, p. 43) We augment categories with keywords extracted from Wikipedia to capture the diverse vocabulary that may reflect various user interests. We also use the topology of categories to explore the diversity of various topics. The last focus explores the last hypothesis by matching page models with the content-based user model derived from Wikipedia, considering serendipity. We compare our approach with the pure content-based model — vector space model — that is introduced in section 2.1.6. In order to test our hypotheses, we have designed the prototype of an agent system to implement our proposed approaches, which are explained in Chapter 3.

Preview of Research Model in Chapter 4

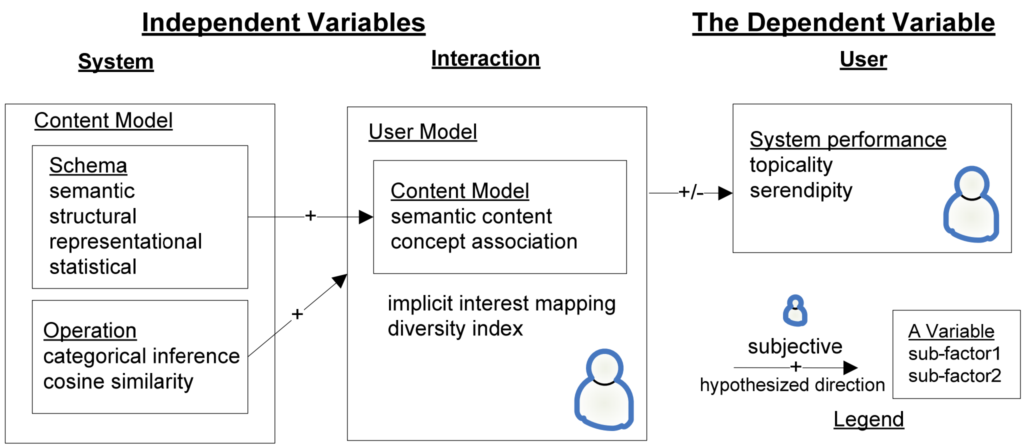

Figure 1.4-2 Visual Model of Research Variables

Figure 1.4-2 displays the visual model of research variables. They are previewed here but details and definitions are explained in section 4.1. Two independent variables –

the content model and user model – are defined. The content model is constructed through different heuristic means (semantic, structural, representational and statistical) and has two operations (categorical inference and cosine similarity). The user model captures implicit interest mappings by calculating and accumulating content models based on page visit frequency. The dependent variable – system performance – is measured through topicality and serendipity. Our research questions are based on the above research model. We evaluate our system's performance using human judgments. The evaluation investigates our main research question QA.

QA. "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*"

Significance

Our work makes four important contributions. First, we emphasize the convergence between content modeling and user modeling as the formulation of a user-derived ontology. This approach could be more dynamic, sharable, unified, and extensible than the traditional vector-based user modeling approaches. We justify the theory of convergence by making use of a knowledge base as the back-end of our recommender system. We also evaluate the system's performance in terms of topical relevance.

Second, it is worthwhile to use Wikipedia's content (categories and keywords) to automatically construct and train a knowledge base for user modeling purposes. To our knowledge, the combination of Wikipedia's categories with heuristic information extraction on keywords for user modeling is a novel attempt. In addition, since we construct the knowledge base through applying information extraction on heuristics, our system can serve as a research platform in which more heuristics may be investigated.

Third, we model a user's topical interests using Wikipedia's categories, which yield a simple model that can be interoperated among different websites. Additionally, our system provides CWP by making use of usage data at the client side, which is a newly-explored research area. Users have control of the client side usage data and thus have dominating influence on recommendations. Furthermore, our platform is capable of adapting its access configuration to address issues like privacy protection or model coverage.

Finally, semantic recommendations bring the awareness of the dynamic topics from Wikipedia to which different mindsets attend. Based on those topics, our system's users can be notified of potentially useful information from any WWW page through recommendations, in addition to intentional searches or random discoveries. Expertise discovery or networking with users of similar interests is another application. Furthermore, we supply a methodology to researchers for further development of similar recommender systems.

Organization of the Rest of this Document

As an overview of the dissertation, Section 2.1 explores different knowledge models. We look into different WP systems in section 2.2. Based on the literature review in Chapter 2, we propose a system with its supporting theories in Chapter 3. We investigate our research questions about the relation between content and user models within the system's context. Involved issues include operational definitions, model alignment, system evaluation, and elements for the recommendation, all of which are explored in Chapter 4. Finally, Chapter 5 covers the future work and potential applications.

Chapter 2 Current Approaches of Web Personalization

User modeling (UM) plays a key role in personalization. We use the definition "A *user model* is a knowledge source in a natural-language dialog system which contains explicit assumptions on all aspects of the user that may be relevant to the dialog behavior of the system. These assumptions must be separable *by the system* from the rest of the system's knowledge" (Wahlster & Kobsa, 1989, p. 6). We analyze UM through the following three dimensions: 1) model acquisition and source, 2) model learning, 3) knowledge representation model.

Model Acquisition and Source

Manual (explicit) and automatic (implicit) are two approaches to formulate initial user models, based on different sources for model acquisition (Quiroga, 1999, p. 16). On the one hand, manual acquisition of user models involves explicit configuration or input from users. Chin phrases it as user-initiated (1993) and the source of model acquisition is direct. On the other hand, the automatic approach (Chin, 1993) often makes use of a knowledge base, e.g., commonsense (Speer, 2007), or machine learning techniques, e.g., association rules (Agrawal & Srikant, 1994), to infer a default user model implicitly. The source for model acquisition is indirect (Chin, 1993).

Model Learning

Modeling learning is similar to model acquisition. Models can be updated explicitly through users' acknowledgement of the system setting, or implicitly through inferring user's behavior (Quiroga, 1999, p. 17; Rich, 1999). Chin phrases it as active interaction versus passive actions that are invisible (1993). The manual/automatic, explicit/implicit, direct/indirect approaches are similar concepts and the terminology often mixed, which is supported by Belkin's suggestions of UM in information retrieval (1997). He advocates that a user model should be formed to share the responsibility of articulating or eliciting the information needs of users. This is often an interactive and iterative process.

Overview of Knowledge Representation Models

Studies use various knowledge models and the implementation to understand a user semantically. We investigate taxonomies, folksonomies, semantic network, ontologies and vector space model in section 2.1, which are relevant to our research scope. Then, as case studies, we review the four major categories of WP systems (Eirinaki & Vazirgiannis, 2003) in section 2.2 – manual decision-rule systems, collaborative-filtering systems, content-based-filtering systems, and web usage mining. Evaluation metrics for modeling performance are reviewed in section 2.2.6. Following these case studies, we discuss the rationales and justifications of our approaches and decisions in section 2.3. Literature about Wikipedia is explored at the end of this chapter as it is our major selection.

2.1. Knowledge Models

2.1.1. Taxonomies

In the human-guided process of deciding how to interrelate information, people place unorganized information into some sort of hierarchy (Bowman, Danzig, Manber, & Schwartz, 1994). Following this intuition, domain experts have traditionally used taxonomies to categorize knowledge domains. Taxonomies are controlled vocabularies from authoritative sources, such as publications, thesaurus, synonyms, or other well-known classification systems, such as Dewey (1876). The use of taxonomies is a top-down knowledge categorization approach. Definite information can be identified systemically given sufficient domain knowledge and a standardized vocabulary. In addition, the hierarchical relation enables semantic knowledge navigation and disambiguation. Taxonomies also provide a means of meta-level reasoning. We can express semantic relationships in taxonomies at various granularities. Furthermore, taxonomies can apply to general knowledge categorization as well as specific domains.

Despite these advantages, taxonomies may fail to reflect sophisticated knowledge comprehensively, due to language ambiguities or differing human perceptions. Furthermore, the mapping or alignment of different taxonomies is sometimes difficult, which consequently lowers its interoperability. According to the IEEE glossary, interoperability is defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged." (Engineers, 1990) Greaves et al. view interoperability from two aspects: syntactic interoperability and semantic interoperability (Greaves, Holmback, & Bradshaw, 2000). In our work, semantic interoperability is the focus. Given the limitations of taxonomies, other approaches have been developed and they are discussed in following paragraphs.

2.1.2. Folksonomies

Vander coined the term "folksonomy" from folk and taxonomy in a mailing list discussion (Smith, 2004). The term delineates the system that people collaboratively assign tags to categorize and describe content, a bottom-up approach as opposed to taxonomies. Folksonomies (Brooks & Montanez, 2006; Mathes, 2004; Quintarelli, 2005) have become popular since the inception of Web 2.0. This approach is intuitive and provides serendipity (Ohkura, Kiyota, & Nakagawa, 2006). Strictly speaking, the format of folksonomies is not a knowledge representation, but an approach to knowledge derivation presented in a tagging format. Although the idea of keyword indexing is not new, the shift from an expert indexer to a "user indexer" is a breakthrough.

Folksonomies is an approach that summarizes the languages a group of users share, considering associative relation in addition to the hierarchical relation that taxonomies focus on. It thus solidifies users' perceptions on information, but produces inconsistencies as well. One major difference between taxonomies and folksonomies is the perspective from which a piece of information is being interpreted. As stated in Brooks and Montanez (2006), there are three basic tagging strategies: 1) annotating information for personal use, 2) placing information into broadly defined categories, and 3) annotating particular articles so as to describe their content. Therefore, folksonomies allow contextual and structural metadata while taxonomies consider the nature or relationship of the information itself. Contextual metadata include author, date of creation, etc., while structural metadata describe the type of a media and its length, etc. In other words, folksonomies focus on information associations as being important as the very content of the information. This orientation draws our focus from the information itself to the perceptions and descriptions of the information. The shift implies that a knowledge space should be constructed by taking perceptual associations or keyword relations into consideration, which brings up the topic of semantic networks.

2.1.3. Semantic Networks (SN)

One research area related to folksonomies with roots in psychology is the semantic network (Quillian, 1968). It allows any kind of relation, such as is-a, relates, has, and thus sometimes is referred to as the Associative Network. Even though there are other similar techniques like Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), the network structure of the SN further provides explanation capability, another unique attribute. In an SN, each node represents a concept. Domain experts predefine the links between nodes. With such an SN framework, associative retrieval is accomplished through the Spreading Activation (SA) technique, which triggers relevant nodes according to a selected activation function.

The SN has the same disadvantage as taxonomies: the need to be constructed by domain experts or even psychologists. Due to the ambiguity of human languages, defining a precise association from an expert's perspective may be different from the general public's perspective. In addition, the interoperability needs to be resolved as well due to the low reusability of the SN. The SN is thus, regarding features, similar but different from folksonomies per se. Furthermore, it requires some tuning of the network structure and activation rules, which is a time-consuming process. Recently, in view of the laborious process, model reusing or exchange has become an alternative. The next topic to be examined addresses this issue.

2.1.4. Ontologies

The definition of ontologies is introduced in section 1.3. “An ontology provides an explicit conceptualization (i.e., meta-information) that describes the semantics of the data” (Fensel, 2003, p. 1), therefore allowing meta-level reasoning. The use of ontologies separates domain knowledge from operational knowledge (Buitelaar, Cimiano, & Magnini, 2005), which is an important concern in model reuse or exchange. Thus, ontologies address the issue of interoperability and semantic representation.

There are plentiful examples of using ontologies in various applications. Castells, Fernandez, and Vallet (2007) use ontologies to improve the performance of information retrieval over large document repositories, or corpora. Davies, Fensel, and Harmelen (2003) use ontologies to manage knowledge particularly for the semantic web environment in their On-To-Knowledge project. Similarly, Maedche, Motik, Stojanovic, Studer, and Volz (2003) primarily focus on the internet environment while Fensel (2003) highlights the power of ontologies for E-commerce. Moreover, there are some specific applications. Guarino, Masolo and Vetere use ontologies to improve content-based search (1999). Dai and Mobasher (2002) use domain/personal ontologies to characterize usage profile and recommend similar items to users. Middleton, Shadbolt, and Roure use ontological inference for user profiling in recommender systems (2004).

2.1.5. Discussion on Knowledge Models

Taxonomies, folksonomies, and the SNs all have strengths and limitations. Taxonomies are systematic but they are controlled by domain experts. Consequently, taxonomies lack flexibility from the user's perspective. Folksonomies aggregate user notations of many kinds of associations for personal uses. The SN uses an expert-constructed network to represent semantic associations. Although semantic associations are intuitive to human information processing, some higher level tasks require the incorporation of taxonomic and ontological resources (Ponzetto & Strube, 2007). An example of higher level tasks is disambiguation in information processing and language understanding. Moreover, Chakrabarti, Dom, Agrawal, and Raghavan have demonstrated that taxonomies improve search and navigation performance in textual databases (1997).

In conclusion, there is a trade-off between systematization and flexibility. Taxonomies and SNs rely heavily on the underlying knowledge structure to form a representation, which to a certain degree is controlled rigidly by experts. However, it is because of the rigidity that taxonomies and the SN are systematic. On the contrary, folksonomies do not have a definite knowledge structure but attempt to formulate one from the user's language. Folksonomies are therefore more flexible than taxonomies and the SN. However, this kind of formation tends to generate a loose knowledge structure. Ontologies capture both advantages of systematization and flexibility in addition to interoperability and conceptualization. Recently, research on ontology learning (Buitelaar, Cimiano, & Magnini, 2005) has started to consider the possibility of constructing an ontological knowledge model from user-generated data, such as blog sites. This development inspires us to consider the prospect of constructing *an ontological representation of the knowledge model from user-generated data.*

The reliability of user-derived ontologies may not be as high as expert-conceived ones but user-derived ontologies may better reflect linguistic perceptions. If reliability or quality is a concern, we can apply ontology derivation from usage data before further validation from experts, which reduces their efforts. In addition, a user-derived ontology may shape the language used by the user or group. Furthermore, ontologies derived from users can be designed deliberately to contain relations that allow knowledge inferences. Moreover, ontologies can act as formal specifications, which eliminate data inconsistency as well. In summary, due to all the mentioned advantages, we selected an ontological model from user-generated web pages for our system.

2.1.6. Vector Space Model: a Content Model in Information Retrieval (IR)

Vector space model (VSM) (Salton, Wong, & Yang, 1975) is the actual implementation of various knowledge models. It abstracts information from web pages using term representation in vector format. Its process includes 1) document preprocessing, 2) term extraction, 3) vector generation, and 4) prediction (Weiss, Indurkhya, Zhang, & Damerau, 2004). The first step eliminates noisy data, such as annotational or navigational texts, and standardizes a document. Second, a text document is tokenized into terms for further processing. Common stemming methods for term cleaning include the truncation of the last few letters, dictionary-based stemming, or the snowball algorithm (Porter, 2001). Each term may weigh differently given a weighting algorithm like term-frequency-inverse-document-frequency (TF-IDF)(Spärck Jones, 1972). TF-IDF values the importance of a term as based on its frequency in a single document and its frequency in the document collection. Terms can also be expanded through various techniques: the use of a thesaurus, synonym word collocation (Lin, 1998), or LSI (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). The expansion handles polymorphs or ambiguities from different contexts. To reduce noisy information, frequently-appearing stop words should be removed if possible.

Third, terms formed in the previous step become vector elements that represent a document. Finally, researchers commonly use cosine similarity (Garcia, 2006) to predict similarity between a new document and others in pairs. As is discussed in section 2.2, many recommendation systems, such as WebWatcher (Joachims, Freitag, & Mitchell, 1997), Syskill & Webert (Pazzani, Muramatsu, & Billsus, 1996), and WebMate (Chen & Sycara, 1998), adopt VSM.

Suggestion for Refining VSM

Unfortunately, VSM often produces a high-dimensional matrix with sparse data. One solution is to use dimension-reduction algorithms to map a high-dimension matrix into a lower one based on support vector machines (Kim, Howland, & Park, 2005). The computational complexity can be highly reduced without losing informative data.

Another solution is to extract only essential information to form key terms. Depending on the area of interest in web content, appropriate information-extraction techniques (Cardie, 1997; Charniak, 1993; Collins, 1999; Manning & Schütze, 1999;

Salton & Buckley, 1988) can be adopted to formulate key terms in the VSM.

Similarly, a clustering or classification process can aggregate similar terms to form a conceptual or semantic VSM. In both Sutcliffe's (1991) and in Wendlandt and Driscoll's (1991) work, they employ experts for concept association and formulating a semantic VSM. Liu (1997) automates the concept association by utilizing heuristic syntax parsing and semantic metrics.

Our Approaches to Content Modeling

Relating to the reviewed knowledge models so far, we combine taxonomies with folksonomies as a user-derived ontology from Wikipedia. VSM is selected as the model implementation. We adopted both auto-extraction and auto-classification approaches for refining VSM and adding semantics. Details are described in section 3.1.2.

2.2. Web Personalization (WP) Systems

In this section, we review WP systems as case studies. WP systems fall into four major categories: 1) decision-rule systems, 2) collaborative-filtering systems, 3) content-based-filtering systems, and 4) usage-based systems (Eirinaki & Vazirgiannis, 2003). As an overview, Table 2-1 lays out the approaches these systems use and gives certain examples of what kinds of knowledge models they commonly adapt. Our approaches are also positioned in the table. For each category of the WP systems, we review certain cases from the literature. Content-based filtering systems and usage-based systems are our focus.

Table 2-1 WP Systems and Approaches

WP Systems ► ▼User Modeling Dimensions	**Manual Decision Rule Systems**	**Collaborative filtering Systems**	**Content-based filtering Systems**	**Web Usage Mining**	**Ours**
Source (direct/indirect)	indirect	direct	indirect	direct	indirect
Acquisition (manual / automatic)	manual	manual, automatic	automatic	automatic	automatic (mostly), manual
Learning (explicit / implicit)	explicit, implicit	explicit, implicit	explicit, implicit	implicit	explicit, implicit
Knowledge Representation (taxonomies/ folksonomies/ semantic network/ ontologies/ VSM/ Others)	stereotypes, semantic network, case-based reasoning, ontologies etc.	tuple (item, ratings), VSM, etc.	VSM, semantic network, ontologies etc.	tuple (item, accesses, time) etc.	semantic VSM, ontologies, taxonomies, folksonomi es

2.2.1. Decision-Rule Systems

Decision-rule systems explicitly formulate user preferences through experts and/or users' interaction or configurations manually. A knowledge base is often the backend platform of these systems. There are various types of knowledge bases. The GRUNDY (Rich, 1979a, , 1979b, , 1999) system is a seminal work that utilizes demographics-based stereotypes integrated with user interaction for book recommendation. The system uses keyword input to derive an initial model. The model is updated explicitly through user feedback. Similarly, Lifestyle Finder

(Krulwich, 1997) uses demographic data to recommend websites, music, purchases, or activities in a question-answer format.

To name some recent works that use different inference mechanisms, Entrée recommends restaurants through case-based reasoning to construct a knowledge base (R. Burke, 2000; R. D. Burke, Hammond, & Young, 1996, , 1997); the Wasabi Personal Shopper (R. Burke, 1999) also utilizes case-based reasoning (Hammond, 1989; Kolodner, 1993; Riesbeck & Schank, 1989) for accessing electronic product catalogs; the Intelligent Labeling Explorer adopts the semantic network to provide labels for exhibits in a museum gallery (O'Donnell, Mellish, Oberlander, & Knott, 2001).

Limitations of Decision-Rule Systems

Manual decision-rule systems rely on collected knowledge and they have limitations. Generally speaking, using pre-defined models, such as stereotypes, cannot represent every user precisely due to the variance of individual differences or situational contexts. The accuracy of the modeling process depends on the underlying knowledge base as well as the interaction process. A knowledge base may not reflect a user's characteristics properly if the user is not a typical member of the target population. In addition, if the user interacts with the modeling system in a different way each time, then the system may have difficulty formulating a consistent user model. Moreover, manual input or configuration is a laborious process. Although certain auto-detected mechanisms (e.g., search caches or domain addresses) can be applied to eliminate the input overhead, there are always implementation constrains to machine-oriented techniques. Furthermore, knowledge bases require the maintenance of experts and system administrators. In view of these limitations, researchers have developed other types of WP.

2.2.2. Collaborative-filtering Systems

Pioneer studies in collaborative-filtering systems make use of subjective user behavior, particularly ratings, in a user community (Goldberg, Nichols, Oki, & Terry, 1992; Herlocker, Konstan, Borchers, & Riedl, 1999; Hill, Stead, Rosenstein, & Furnas, 1995; Kautz, Selman, & Shah, 1997; Konstan et al., 1997; Shardanand & Maes, 1995). The main idea assumes that people who have similar ratings for commonly rated items will have similar ratings for unseen items. Its approach for model formulation is opposed to manual decision rule systems that require explicit user input as direct knowledge acquisition. Collaborative-filtering imposes parts of the knowledge acquisition process on a user group through an indirect input, which derives models from peer behavior and product ratings or selections, rather than predefined domain knowledge. Web usage mining techniques, which will be discussed shortly, often incorporate collaborative filters since user ratings or content are treated as usage data. Nevertheless, the rationale of collaborative-filtering systems is essentially the same as stereotypes because of the assumption that users who share common interests will rate specific items similarly.

However, collaborative-filtering systems suffer the problem of insufficient ratings or usage on every available item as well as newly added items (Mobasher, Dai, Luo, Sun, & Zhu, 2000). This is commonly referred as the cold start problem. Another common phenomenon in collaborative-filtering systems is Price's Cumulative Advantage Model (Price, 1976). The model explains that popular articles always stay at the highest ranks, which suggests that success breeds success. In any case, collaborative-filtering systems are not the primary focus of this dissertation, but we introduce them due to their relevance to usage mining. Readers who are interested in collaborative-filtering can refer to Herlocker, Konstan, Terveen, and Riedl's study (2004) or the recent survey paper (Schafer, Dan, Jon, & Shilad, 2007).

2.2.3. Content-based-filtering Systems

Content-based-filtering systems track a user's preferences and generate recommendations that are similar to the user's past selections. These systems make recommendations using IR approaches, particularly the Vector Space Model (VSM), and various machine learning techniques. Techniques include TF-IDF, *k*-nearest neighbor, reinforcement learning, self-organization maps, decision trees, naive Bayes, neural networks, and support vector machines. The study (Michael & Daniel, 2007) provides an overview of current methods and strategies for content-based recommendation systems. In this section, we describe three classical systems that recommend pages across websites in details -- WebWatcher (Joachims, Freitag, & Mitchell, 1997), Syskill & Webert (Pazzani, Muramatsu, & Billsus, 1996), and WebMate (Chen & Sycara, 1998). There is a discussion of the pros and cons of each approach at the end.

WebWatcher (Joachims, Freitag, & Mitchell, 1997)

WebWatcher works as a proxy to observe clicking behavior and recommend interesting web links displayed at a currently browsed page. Keywords and structural mining on pages are the major modeling approach. The system initially asks users to input keywords that describe their interests as an explicit profile. The state of a user profile is updated as a user clicks on links. WebWatcher augments the user profile by adding the clicked anchor text from html tags, which illustrates the importance of anchor texts. The system adopts VSM as the user profile and page model. A current user profile is then compared with a vector set computed from web pages downstream of the currently-browsed webpage. This comparison is achieved through learning from the hypertext structure using reinforcement learning. It allows the system to maximize the relevant information encountered over the trajectory of possible browsing paths. This is essentially the same concept as the information foraging theory that was discussed in section 1.2.

WebWatcher examines every link in the currently-browsed page and then pre-processes forthcoming links belonging to each examined link respectively. WebWatcher defines an evaluation function to derive the optimal link(s) that have the maximum similarity value with the current user profile. The similarity between the user profile vector and each vector in a set of web pages is calculatedas couples sequentially using cosine similarity measure. WebWatcher recommends three optimal web links via highlighting.

Syskill & Webert (Pazzani, Muramatsu, & Billsus, 1996)

Syskill & Webert recommends arbitrary pages in the World Wide Web according to user ratings in three degrees (Hot, Luke-warm, and Cold). The system constructs a multi-topic profile for each user. A topic can be as broad as “movies” or as specific as “protein.” For any topic in the predefined collection, Syskill & Webert initially asks a user to rate pages in a predefined “index page” collection. Based on the ratings, the system either begins to make recommendations on pages accessible from index pages, or queries the user's interests on LYCOS (Maudlin & Leavitt, 1994) for more information in preparation for the next rating. The interests of a user are re-formulated each time feedback is obtained. Syskill & Webert annotates the list of page links as recommendations, which suggest links a user should and should not click.

The underlying recommendation techniques of Syskill & Webert relies on a Bayesian classifier, which is elaborated in studies like Duda, Hart and Stork's (2000). This classifier was selected after running an empirical comparison among five other classifiers (Nearest Neighbor, PEBLS (Cost & Salzberg, 1993), Decision Trees, TF-IDF, neural network). A vector representing a user's interests is formed through feature extraction from the combination of highly-rated pages. A feature can be a single word or a phrase based on the calculation of expected information gain (Quinlan, 1986) from those pages. Expected information gain identifies informative keywords according to the presence of keywords in the Hot list and the absence of keyword in the Cold list. Any web page potentially of interest to the user is transformed into the vector representation, and the probability of interest is calculated through the Bayesian classifier.

WebMate (Chen & Sycara, 1998)

Functionally speaking, WebMate is analogous to Syskill & Webert in that relevant web pages are recommended based on keyword modeling and relevance feedback. In addition to providing navigational recommendations, search refinement is advised as well. The underlying techniques of WebWatcher are conceptually equivalent to many of those introduced in this section: VSM model, TF-IDF weighting, cosine similarity Measurement. However, instead of applying a classifier based on categories in topical domains, WebMate clusters vectors with high similarity according to a predefined threshold. Another major difference is that WebMate uses the Trigger Pairs Model (Gauch & Futrelle, 1993; Rosenfeld, 1994), a combination of statistical and linguistic features, to augment search keywords; The model defines a “trigger pair” as a word *S* that is significantly correlated with, and thus triggers, another word *T*. Mutual Information that considers the word order is the core measurement of this model. Mutual information quantifies the two variables’ dependency.

Pros and Cons of Content-based Filtering Systems

These three selected systems present a content based approach to model and assist users across different websites. Similar to GRUNDY, Syskill & Webert and WebMate adopt a dialogue-like interaction for user feedback. This kind of explicit modeling is in contrast to the implicit modeling constructed through clicking behavior in WebWatcher (Joachims, Freitag, & Mitchell, 1997). Models formulated in the content based approach are ideal for cross-website personalization because contents are highly accessible. Nevertheless, as a major challenge of personalization systems, these models are either application or service dependent, which makes them difficult for another personalization system to reuse.

There are other pros and cons of content-based-filtering systems. Generally speaking, implicit content-based filters are useful for pre-evaluating newly-added items prior to subjective judgments. Relevance feedback further facilitates learning a user’s interests. Nevertheless, there are certain disadvantages in those systems that utilize the aforementioned IR techniques. As pointed out by Balabanovi and Shoham, content-based recommendation systems “completely ignore aesthetic qualities, all

multimedia information, and network factors such as loading time (1997)."
Over-specialization (over-fitting) is another problem in systems like Syskill & Webert WebWatcher and WebMate. It occurs when users can only receive recommendations similar to those previously highly-rated pages (Balabanovi & Shoham, 1997). Due to these shortcomings, content-based-filtering is often combined with web usage mining, which is the approach we choose.

2.2.4. **Web Usage Mining**

Web usage mining (Agrawal & Srikant, 1994; Eirinaki & Vazirgiannis, 2003; Mobasher, Cooley, & Srivastava, 2000) studies the patterns and relationship between web logs and web users, particularly from a single website. It is an unobtrusive way to formulate an implicit user model. The term "web log" in the context of usage mining is referred to as any server-side recorded data, including web servers and database servers, etc. One common example is the log files, which contain the site visits, addresses, search keywords, and timestamps when pages or files are retrieved from a web server.

The procedure of web usage mining includes 1) data preprocessing, 2) pattern discovery, and 3) data analysis (Srivastava, Cooley, Deshpande, & Tan, 2000). Firstly, during the data preprocessing phase, data analysts cooperate with domain knowledge experts or analysts to clean data and remove irrelevant information, handle missing values and ensure data correctness and consistency. Usually, a standardized abstraction of the cleaned data is formed in preparation for comparison.

In the second phase, pattern discovery detects patterns within data through the following techniques: statistical analysis, association rules, clustering, classification and sequential analysis (Eirinaki & Vazirgiannis, 2003). Statistical analyses study examples like page access activities, search engine referral or keyword search activities, and user demographics. While the majority of these pattern discovery techniques are self-explanatory, association rules and sequential analysis require further clarification. Association rules (Agrawal & Srikant, 1994) infer the relationship between web pages through users' logged behavior. Sequential analysis (Mannila, Toivonen, & Verkamo, 1995) discovers frequent sequences, with a temporal relation, and patterns among web page access behaviors. Association rules and sequential analysis are often combined with clustering or classification techniques to optimize results of the discovered patterns.

Third, once patterns have been discovered, data analysts invite domain experts to interpret the patterns in the analysis phase. Domain experts then provide appropriate modifications or system improvements. Web usage mining may be an iterative learning process depending on the technique(s) used and the nature of data.

Content-Based Web Usage Mining

The techniques of web content and usage mining (Balabanovi & Shoham, 1997; Eirinaki, Vazirgiannis, & Varlamis, 2003; Mobasher, Dai, Luo, Sun, & Zhu, 2000) complement each other, suggesting a hybrid approach. Web usage mining has a similar problem of insufficient usage for newly-added items as collaborative-filtering systems have. In addition, web usage is highly related to the structure and layout of web pages. As adaptive hypermedia become popular, more and more websites generate their content or even structure dynamically. Under this circumstance, the identification of the access patterns of web pages can become less informative unless the semantic content of usage data is taken into account. To tackle these problems, research has shown that integrating content into web usage mining can yield more effective WP (Mobasher, Dai, Luo, Sun, & Zhu, 2000).

Finally, most of the above mentioned processes are applicable to the scenario of an individual user's log data from multiple websites, particularly when the semantic perspective is involved. The pattern to be discovered is a user's persistent interest or habitual browsing behavior. It should be noted that the characteristics of collected data and the selections of analytical approaches are important considerations.

2.2.5. Practical Guidance

We briefly summarize characteristics for WP using Kobsa's overview of practical systems. Kobsa analyzes existing user modeling systems and lists the following required services and characteristics frequently found in a modeling system:
Services (Kobsa, 2001, p. 56) :

1. The representation of assumptions about one or more types of user characteristics in models of individual users (e.g. assumptions about their knowledge, misconceptions, goals, plans, preferences, tasks, and abilities);
2. The representation of relevant common characteristics of users pertaining to specific user subgroups of the application system (the so-called stereotypes);

3. The classification of users as belonging to one or more of these subgroups, and the integration of the typical characteristics of these subgroups into the current individual user model;
4. The recording of users' behavior, particularly their past interaction with the system;
5. The formation of assumptions about the user based on the interaction history;
6. The generalization of the interaction histories of many users into stereotypes;
7. The drawing of additional assumptions about the current user based on initial ones;
8. Consistency maintenance in the user model;
9. The provision of the current assumptions about the user, as well as justifications for these assumptions;
10. The evaluation of the entries in the current user model, and the comparison with given standards. (Kobsa, 1995, pp. iii-v)

Characteristics (Kobsa, 2001, pp. 57-59):

11. Generality, including domain independence
12. Expressiveness and strong inferential capabilities
13. Support for quick adaptation
14. Extensibility
15. Import of external user-related information
16. Management of distributed information
17. Support for open standards
18. Load balancing

19. Failover strategies

20. Transactional consistency

21. Privacy support

Discussion on User Modeling Guidance

To summarize the literature review about user modeling, there are five points for discussion. First, users should be modeled individually so that web services can be better personalized. Although many of the above listed systems address stereotypes, additional implications need to be inferred beyond general user interests, as stated in the 7^{th} item in the service list.

Second, assumptions and common characteristics of users should be captured at different granularity levels, specifically at the motivational, motional (action), and operational level. For example, in the web environment, we assume that users with a definite goal generally perform a search, while browsing may imply an indefinite goal. This is a common assumption that shapes users at the motivational level of abstraction (i.e. definite vs. indefinite) while the corresponding action is search and browsing. The next example is a detailed assumption of users at the operational level of abstraction: an URL click after performing a keyword search implies that the displayed text in the search result perceptually matches the user's search goal. These reflections of users at various granularities are crucial in order to identify what kinds of interaction data to collect and which analytical techniques or approaches to apply. Kobsa's guideline is simply a very general suggestion that needs to be tailored to the web environment, which is one of our goals for this work.

Third, mining and retrieval techniques help us to construct a knowledge base in which areas of interest (e.g., behavioral patterns) are focused. In addition to user models, sophisticated knowledge representation is often necessary when the semantic level is considered as well.

Fourth, mapping usage data to the knowledge base reveals insight into a user's interests, intentions, or even relationship information. Although personalization systems can ask for user input to provide better customization, user effort should be minimized whenever possible. Users should only be prompted for necessary relevance feedback, such as handling the *over-specialization* problem mentioned in the discussion of content-based-filtering agents.

Fifth, user models should be maintained and evaluated constantly, as items 8 to 10 in the above service list and the design decisions of many aforementioned systems in section 2.2 suggest. In addition, domain independence is an important concern in model adaptation or extension. Furthermore, interoperability of user models should be ensured through mechanisms, such as importing and exporting protocols or standards, such as OWL (W3C, 2004).

2.2.6. Performance Evaluation

Recall and precision are standard metrics in IR (Baeza-Yates & Ribeiro-Neto, 1999). Recall reveals the fraction of relevant documents that a prediction identifies and retrieves in a document collection; precision measures the proportion of relevant documents to all retrieved documents in the document collection. There tend to be a trade-off between recall and precision in practice, i.e., high recall implies low precision or vice versa.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Although these metrics are quantitative, relevance assessments include other factors that are subjective, dynamic, cognitive, situational, multidimensional and context dependent (Quiroga & Mostafa, 2002; Schamber, 1994). Examples are motivation, accessibility, reliability and authorship, etc. Additionally, other factors, such as novelty or serendipity of recommendations, influence user satisfaction in evaluating recommendations (Herlocker, Konstan, Terveen, & Riedl, 2004). While no system can achieve perfect relevance, how to define high relevance is still a matter of research.

In our evaluation, we assess the relevance of recommendations focusing mainly on two variables: topicality, and serendipity. We particularly address the diversity of recommendations in order to promote serendipity. Details and definitions are discussed in Chapter 4, particularly in section 4.2.4 and 4.3.

2.2.7. Diversity and Serendipity

Related work on promoting diversity in recommendation includes Symth's discussion regarding the tradeoff between similarity and diversity (Smyth & McClave, 2001) and Ziegler's effort in looking into intra similarity among different recommendation lists (Ziegler, Lausen, & Schmidt-Thieme, 2004). Both studies are about the diversity of a selected topic, which differs from our focus on the diversity of a user's interests in different topics within a domain. The identification of a user's interests and interest coverage, defined in section 3.1.5, is important as recommending pages across topics or websites may go beyond content similarity to diversity or serendipity.

A concept related to serendipity is information encountering (Erdelez, 1997). A broad definition of information encountering is a memorable experience of gaining valuable or interesting information out of a random exploration. Other researchers use a different terminology to express a similar idea: incidental information acquisition (Williamson, 2002), opportunistic acquisition of information (Erdelez, 1997), serendipitous information retrieval (Toms, 2000), or opportunistic communication (Budzik, Bradshaw, Fu, & Hammond, 2002). Erdelez (1997) proposes three key elements of information encountering: 1) characteristics of information user, 2) characteristics of the information environment, and 3) characteristics of the encountered information. These three elements echo our research foci of semantic user modeling and content modeling in the web environment. We discuss them more in the next section.

2.3. Discussion of Model Alignment – User and Content Models

Research Foci

We assert that the alignment between user and content models addresses our three research foci of content modeling, user modeling, and the match between the two. In section 2.1, selected models for our first focus "content modeling for web pages" have been reviewed. We also have identified the ontological representation inferred from user-generated data that may reflect users' languages in the knowledge domain explored by the user group in section 2.1.5. The representation is thus a potential indicator for capturing users' interests. Section 2.2 reviews certain WP systems and examines our second and third foci “modeling a web user’s topical interests” and “matching page models with a user model.” Then we have drawn the conclusion that mapping usage data into a user model via a knowledge base can reveal insights into the user’s interest in a topical domain. Individual web user’s usage data relevant to our research focus are those pages manipulated (e.g., keywords searched, browsed, or bookmarked) by the user. Within our research scope, we consider a number of associations between the user and usage data, such as access frequency, semantic content, temporal (sequential, ephemeral, and persistent) context, topical diversity, serendipity and situational scenarios etc.

Relevance

Although topical relevance may be treated as an objective factor, there are definitely other subjective variables that affect relevance judgment. The following list summarizes common characteristics of relevance found in the literature (Saracevic, 1970, , 1975; Schamber, 1994; Stephen, 1992):

- Subjective, depending on human judgment and thus not an inherent characteristic of information or document;
- Cognitive, depending ultimately on human knowledge and perceptions;
- Situational, relating to individual users’ information problems;
- Multidimensional, influenced by many factors;
- Dynamic, constantly changing over time; and
- Measurable, observable at a single point in time.

Because of the subjective factors listed in most of the bullets, it is difficult for personalization systems to achieve perfect accuracy. For example, a previously-computed recommendation due to a high access frequency is not necessarily a good recommendation since the user may have had a lot of knowledge in that subject domain. He or she may look for something that interests him or her, probably in other unfamiliar knowledge domains. Chin's work (2007) emphasizes a similar relationship between a user's expertise and the appropriate filtering mechanism. Therefore, although there are quantitative formulas to measure the relevance of interestingness, subjective relevance feedback is unavoidable.

Knowledge Base

Turning back to the discussion of knowledge base, we need to conceive the space carefully with an appropriate knowledge representation. World knowledge is very sophisticated and it is difficult to design a data structure that is representative of the nature or complexity of many knowledge domains. For instance, the taxonomy-like data structure has been criticized for its hierarchical rigidness in modeling complex knowledge. Knowledge sometimes cannot be categorized definitively due to different human perceptions or cognitions. Regarding semantic networks, the network-like structure relies heavily on the weighting or definitions of those links among connected nodes in the network. Quantitative weights and the network structure are difficult to convert into precise explanations for qualitative analysis. Weights could be an ordinal reference but not an interval variable. For example, we cannot say that the semantic relation is twice as strong for a link with a weight 2 as another link with a weight 1. The Tuple-like data structure has a similar problem in that a value scale is used. The lack of precise explanations cannot be avoided unless the semantic level is considered. In addition, the flat nature of tuple-like structure makes it less efficient or less informative to capture the relation between each tuple.

As mentioned, ontologies are flexible and systematic. We can select the domain of interest to formulate an ontological space and add more relations into the space as needed. In addition, Middleton et al. (2004) have shown that using an ontology in user profiling results in superior performance over a flat data structure. Ontologies can also be used to track users at a conceptual level, taking into consideration semantic content (Oberle, Berendt, Hotho, & Gonzalez, 2003) or usage metadata (Dai & Mobasher, 2002). Therefore, we can model users with an ontology that covers a focused knowledge domain, which aligns user modeling with content modeling to a certain degree.

The potential of a user-derived ontology has been illustrated. Instead of explicitly constructing the ontology, an alternative is to adopt a representative online community. We can then use its content and usage data to formulate a community-derived ontology. In this way, a public community website enables us to define an ontological space via modeling web pages from the site.

Choosing Wikipedia

Modeling web users via page content implies adaptability and flexibility since pages are easily accessed and conveniently transferred. People can readily share their languages using the common ground of a unified platform in a public website. This means that the ontology derived from such a site is dynamic per se. In addition, extensibility can be achieved through integrating new content models, on top of an existing one, from web pages in different service domains. An implication of community-derived models is that there may be fewer knowledge experts required in building the content and user models. It is reasonable to speculate that there may be certain experts in a community, who in some way participate in the knowledge (re)formulation as in the case of Wikipedia. Therefore, knowledge experts may analyze and interpret the underlying behavior of a processed user's model, which may save their time.

In conclusion, we need a platform that represents user language at different granularities, such as cultural contexts, personal values, habitual behavior of information acquisition, preferred keywords, and conscious data input, etc. We selected Wikipedia to formulate the community-derived ontology. The collaborative content of Wikipedia may include popular and dynamic topics and vocabularies. It is worth trying to use the derived ontology from Wikipedia as a standardized knowledge base, primarily for the purposes of semantic user modeling considering diversity and serendipity.

2.4. Wikipedia

Wikipedia, one of the world's largest collaborative knowledge bases, was chosen as the platform from which to construct a community-derived ontology. Most web users know Wikipedia due to its high visibility from major search engines. Alexa.com, a leading WWW traffic analysis company, ranks Wikipedia as one of the top 7 websites as of July, 2010. Although there are only a few contributors (less than 10% of all users) to the content of Wikipedia (Ortega, Gonzalez-Barahona, & Robles, 2008; Priedhorsky et al., 2007), it has a huge pool of readers due to its high accessibility. It is therefore assumed that Wikipedia's users, which include authors, readers, editors, reviewers, and administrators etc., can more or less represent or influence general web users. This point is further elaborated more two paragraphs later.

Pros of Wikipedia

There are five justifications to support the selection of Wikipedia based on its features in terms of values for research: 1) quality, 2) diversity, 3) associations, 4) dynamics, and 5) convergence. 1) First of all, a comparison regarding content quality between Wikipedia and *Britannica* has been made and reported in journal *Nature*. The study found a similar level of quality for both encyclopedias, which reveals the potential of Wikipedia (Giles, 2005). Prior literature (Strube & Ponzetto, 2006) also indicated that Wikipedia provides a suitable encyclopedic knowledge base for extracting semantic information due to the potential of collaborative work and the taxonomy structure.

2) With respect to diversity, as Wikipedia has many editors and reviewers with various levels of expertise, it is reasonable to speculate that editors and reviewers enrich the terminology of its content. As Sussan describes, "with Web 2.0 products, it is the user's engagement with the website that literally drives it."(Sussan, 2007, p. 35) Similarly, we speculate Wikipedia's content and its vocabulary may cover recent and popular topical areas that people are generally interested in. The language in Wikipedia may be closer to what the general public uses, instead of being controlled by domain experts. We emphasize that the topics, but not the content accuracy, of Wikipedia may reflect the dynamic nature of information on the Internet.

3) As far as associations are concerned, Wikipedia includes the thematic association of keywords as well as hierarchical classification of topical keywords. Thematic

association is any kind of topical connection that one can conceive of. One example of a thematic association is the “See Also” section in a Wikipedia page. Hierarchical classification is the categorical system created by Wikipedia users[2]. Both the association and the classification are dynamically maintained by community members.

4) In addition to these semantic advantages, the dynamic nature of Wikipedia makes it possible to find a newly coined terms or keyword of significant events that happened recently. This information may not be as rapidly captured by manually-constructed knowledge bases from human indexers or experts.

5) As for convergence, different types of semantic keywords (e.g. associative, categorical, and temporal) plus hypertexts in Wikipedia may evolve to a shared stable condition given a longer time frame. This is due to a continuous review process. Similar to this procedure, researchers have begun to look into the convergence of metadata vocabulary within a community, i.e. keyword assignment for articles, supported by user created keywords (Joseph, Yukawa, Suthers, & Harada, 2006). Joseph et al. found that associative browsing, supported by metadata, is as popular as search. Therefore, given the justifications previously stated, we are inspired to choose Wikipedia to analyze keyword relations.

Cons of Wikipedia

Nevertheless, there are a few general downsides to Wikipedia. Denning, Horning,

[2] http://en.wikipedia.org/wiki/Wikipedia:Categorization

Parnes and Weinstein points out six perspectives on Wikipedia's potential risks (2005). They are listed in no particular order. “1) accuracy, 2) motives, 3) uncertain expertise (the qualification of an article's contributors), 4) volatility (contributions and corrections may be negated by future contributors), 5) coverage, and 6) sources.” These risks are mainly due to the collaborative nature of Wikipedia and the absence of centralized control. Another obvious risk of Wikipedia is that a large body of its content is defined in what Pitkow et al. refer to as “consensus relevance and author relevance (2002).” Consensus relevance presumes that what an entire population deems as relevant is relevant to an individual user as well. This kind of social relevancy may not always imply content (semantic) relevance, considering the accuracy per se, let alone other types of relevance. Author relevance stresses that the meaning and resource values contributed by a group of authors deeply influence the outcome of the entire user population.

Discussion on Wikipedia and Our Decisions

Despite the above downsides, Wikipedia is indeed a platform for knowledge formulation. Wikipedia’s knowledge evolution is drawn by the contribution of enthusiasts. Its high accessibility implies the transparency of the evolution process. Consensus, author and content relevance are all by-products that help the process of achieving correct knowledge before we can build on it. Drawing the attention of people who have interests in the knowledge formulation process is thus important. Therefore, Wikipedia is valuable for our purpose to identify semantically relevant information to the user’s interest.

In conclusion, although Wikipedia raises many controversies, such as accuracy and unequal contributions, the positive side of Wikipedia outweighs the negative side in terms of this research. Consequently, it is a good test bed for building a standardized knowledge base. We use this base for semantic modeling to filter information found in heterogeneous web pages for semantic modeling. Additionally, building on this standardized base, we model users and their interests in certain knowledge domains by means of individual usage data at the client side.

Regarding other work that uses Wikipedia to construct a knowledge base, Yago (Suchanek, Kasneci, & Weikum, 2007) and DBpedia (Bizer et al., 2009) are two notable examples. Yago merges extracted entities and relationships from Wikipedia with WordNet as an extendable ontology with near-human accuracy. DBpedia extracts the structural information from Wikipedia as a database for query and data linking. Both Yago and DBpedia are primarily for dictionary or look-up references, but not for user modeling or personalization. Our knowledge base derived from Wikipedia focuses on semantic user modeling using categories and extracted keywords with weights from Wikipedia.

Another similar attempt to elicit user interests using Wikipedia is found in Szomszor et al's work (Szomszor, Alani, Cantador, O'Hara, & Shadbolt, 2008). They adapt Wikipedia's categories to consolidate user profiles, particularly tags, from two social networking sites: Delicious and Flickr. One major distinction is that our work takes the content-based approach while folksonomy analysis is Szomszor et al's primary effort. The content-based approach applies to any website, so does not limit us to social websites that rely on collaborative filtering or tagging. Furthermore, our work emphasizes examining the coverage of a user's interests, while interest acquisition is Szomszor et al's focus. In the next chapter, our system is described in detail.

Chapter 3 Proposed Process, Model and System

Semantic user modeling across websites is the essential idea of our research. Therefore, our proposed system is within the framework of Cross-System Personalization (CSP) and addresses our research foci of the alignment of content and user model, privacy concerns, topical diversity and serendipity for recommendations. CSP aims to provide personalization based on shared profiles and protocols among different service systems (Mehta, Niederée, & Stewart, 2005). As mentioned in section 1.2, there are two identifiable deficiencies in existing CSP approaches (Zhang, Song, & Zhang, 2006). One deficiency is model replacement from one user-adaptive system to another. The other is the lack of agreement on a shared communication protocol among agents or personalization systems.

Regarding the first deficiency of model replacement, we propose having a unified model at the client-side. As long as there is a shared protocol for model exchange, service providers can utilize the user model for product recommendations. There are three benefits of client-side modeling. First, users have more control over their preferences and privacy if their profiles are generated locally (Kobsa, 2007). Second, client-side usage data is an alternative source for commercial websites to integrate profiles from other websites that are difficult to acquire due to their conflicts of interest. Third, as pointed out by Padmanabhan et al., by modeling usage data from multiple websites, user behavior can be predicted with greater accuracy (Padmanabhan, Zheng, & Kimbrough, 2001).

As for the second deficiency of a shared communication protocol, we propose using ontologies as a standard to addresses semantic interoperability (Greaves, Holmback, & Bradshaw, 2000). The use of ontologies is a key technique that characterizes the semantics of information exchange (Fensel, 2003, p. 4). Nevertheless, constructing or mapping an ontology that accommodates various topics from different systems is labor intensive and hard to maintain. To automate the process, we have selected *Wikipedia*, one of the world's largest collaborative knowledge bases, as the shared platform for ontology derivation.

Our proposed solution not only deals with both deficiencies of CSP, but also addresses semantic user modeling through collaborative content and individual usage data. We intend to test our hypothesis of the alignment between content modeling and user modeling regarding semantic relevance addressing serendipity. This chapter describes the prototype of our agent system that implements the proposed solution. We begin with the overall structure and explain each component in the sub-sections.

3.1. The iPK (my Personalized Knowledge) System

We create the iPK system, which stands for my personalized knowledge from the Internet. iPK mainly makes use of web usage data to formulate the user model. Optional user feedback is obtained to validate the model. Figure 3.1-1 displays the state diagram of the iPK system. There are six major states. We begin with the introduction of "identifying usage data". However, details of the underlying components and techniques among those states are discussed in section 3.1.1 to 3.1.5.

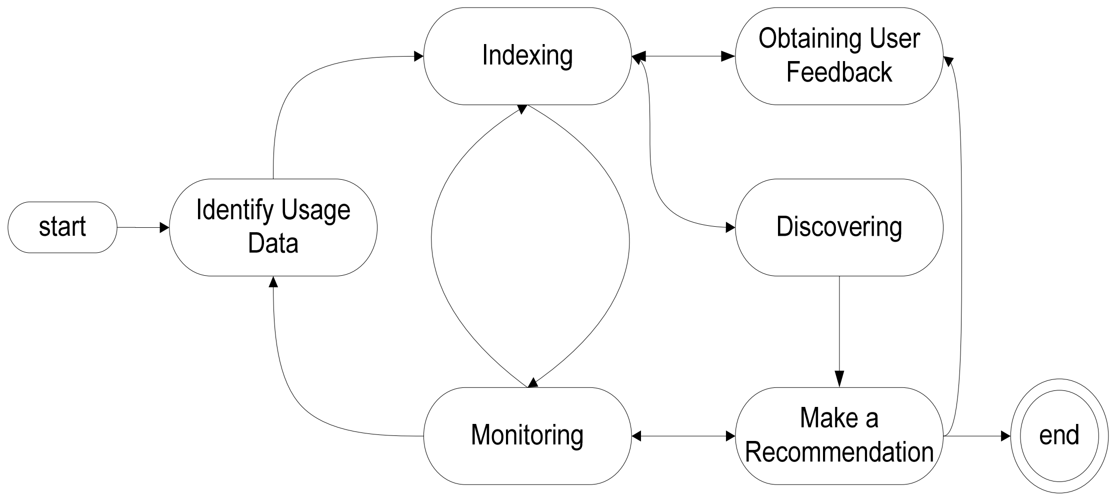

Figure 3.1-1 State Diagram for iPK System

Identifying Usage Data

Given privacy concerns, although technically the system is able to autonomously collect data by itself, we want the user to be aware of what kind of data the system is collecting. Therefore, in the current implementation, the user is asked to locate his or her usage data initially. Examples of usage data are browsed web pages, the page visit log from web browsers, or bookmarks. Then the iPK system can move to the indexing state.

Indexing and Feedback (Explicit Input)

In the indexing state, the system analyzes usage pages and user specified pages in preparation for identifying certain web pages that may require monitoring. It maps usage data into the knowledge base built in the system. The mapped user model is modified in other states and used for discovering new information from the Internet.

Monitoring

The monitoring state includes explicit and implicit monitoring. For explicit monitoring, iPK learns a user's interests from the usage data and creates a "watch list" of frequently accessed pages. Once sufficient knowledge (i.e. one month of usage data according to our experimental findings) is accumulated in the system, iPK alerts the user to review the watch list. iPK regularly checks for updates of those user-confirmed pages in the watch list. As for implicit monitoring, iPK records a user's clicking behavior of recommended pages, which implies the user's interests in the pages. Implicit monitoring is helpful when direct user feedback is unavailable.

Discovering

In the discovering state, iPK discovers and fetches pages from the Internet by employing a crawler and comparing those pages with the user model. The system primarily filters keyword query results from a popular search engine, namely Google.

Recommending and Feedback (Implicit Learning)

The matcher component makes a recommendation, which triggers the recommendation state, whenever the discovered web pages are similar to a user's model and they have not been accessed previously. Recommended pages are monitored implicitly for user clicks or be requested explicitly for user feedback.

3.1.1.System Architecture

Figure 3.1.1-1 below describes the overall architecture of the iPK system. We describe the component details in the subsections of 3.1 and the variable definitions in section 4.1. Basically, there are four major components – knowledge base (WikiBase), sensor, crawler and matcher. Relating to our research contexts, we construct the knowledge base to test the alignment of content models and user models. The sensor generates different configurations, introduced later, of the content model variable. It is also in charge of the updates of the user model. The matcher handles certain other factors, such as relevance predictions, of the user model variable. The crawler explores new pages from WWW.

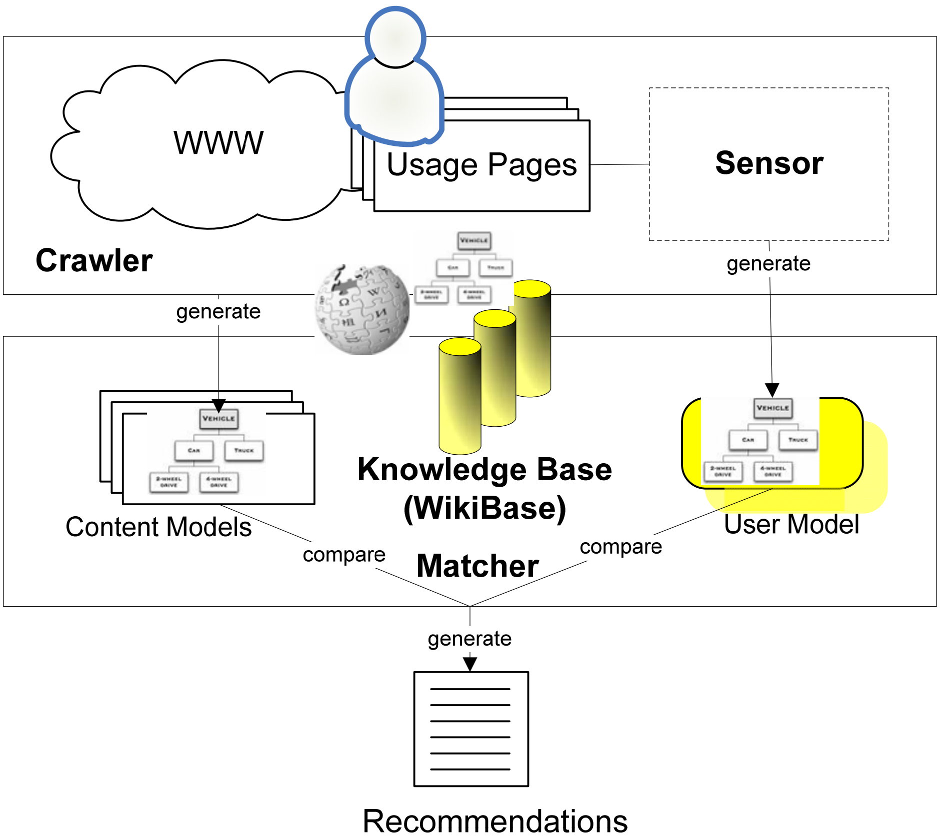

Figure 3.1.1-1 System Architecture

Functionally speaking, the knowledge base stores the community-derived ontology from Wikipedia as well as the user's usage data and models. The sensor performs the acquisition of usage data as well as usage analysis and indexing, which (re)formulates a user model. The user model represents a user's persistent long term interests. The crawler monitors web pages in the watch list and discovers potentially interesting web pages, mainly through linkage mining. Interestingness is defined as the degree of similarity between any web page and the user model. The matcher compares those discovered pages from the crawler with the user model and generates recommendations if they are similar. Any update from the watch list also yields an alert recommendation. Recommendations are presented in digest format, which lists new recommended pages (ranked according to similarity) and updated pages separately. User feedback and a user's configurations of preferred types of recommendations are also managed by the matcher.

3.1.2.Knowledge Base (WikiBase)

There are two major types of data stored in WikiBase. One is the Wikipedia-derived ontology while the other is the system representation of an individual's usage pages as content and user models. For the convenience of reference, the derived ontology is named Wikipedia Ontology (WO) hereinafter. We discuss WO in this section. Content and user models are explained in the next section (3.1.3), as they are managed by the sensor component.

As an overview, we construct WO by augmenting Wikipedia's categories with heuristic information extraction to obtain keywords from pages belonging to the same category. Heuristics include page titles, categorical labels, anchor texts, italic and bold terms, as well as terms with high TF-IDF scores. These keywords are extracted as a collection for each category in Wikipedia. Each keyword has a significance weight that the system utilizes for modeling web pages. The weight is assigned based on the frequency of keywords that appear among all pages belonging to the same category.

We begin with the justifications for augmenting categories with heuristic keywords, followed by the explanations of heuristics. The keyword weighting mechanism is explained at the end.

Justifying the Augmentation

There are three major reasons for the construction of WO. First, the strategy of keyword representation is adapted from folksonomies, which have the characteristics of being ordinary and providing serendipity (Ohkura, Kiyota, & Nakagawa, 2006). People tend to use different terms for a similar concept (Furnas, Landauer, Gomez, & Dumais, 1987), which yields fragments of information. One of our goals is to compile these pieces of information together using Wikipedia. Second, the categorical structure expresses a hierarchal relation and allows for concept tracking or language disambiguation. Third, due to the neighboring relation of the categories, the ontological structure reflects the rationale of conceptual associations or a thematic relation. In this way, the knowledge model of WO is constructed using a network structure based on a conceptual space enriched by contextual keywords. Maintaining both hierarchal and associative relations is intended to combine the advantages of taxonomies, folksonomies and semantic network.

Computer Science Categories

In order to automate the process to construct WO, we need to understand the underlying data source of Wikipedia. Experiments were conducted in the domain of computer science (CS) for in-depth analysis, due to its relatively rich categories and pages in Wikipedia. We selected the root category "computer science"[3] and crawled

---

[3] http://en.wikipedia.org/wiki/Category:Computer_science

out its subcategories up to depth 2, which provided 378 categories[4]. To better control the evaluation process and not to overload participants with too many categories for their judgments, we constructed only two levels of categories.

Figure 3.1.2-1 visualizes the categorical structure of these categories, up to depth 2, using prefuse (Heer, Card, & Landay, 2005). CS is at the central root and the blue nodes are the first level subcategories of CS.

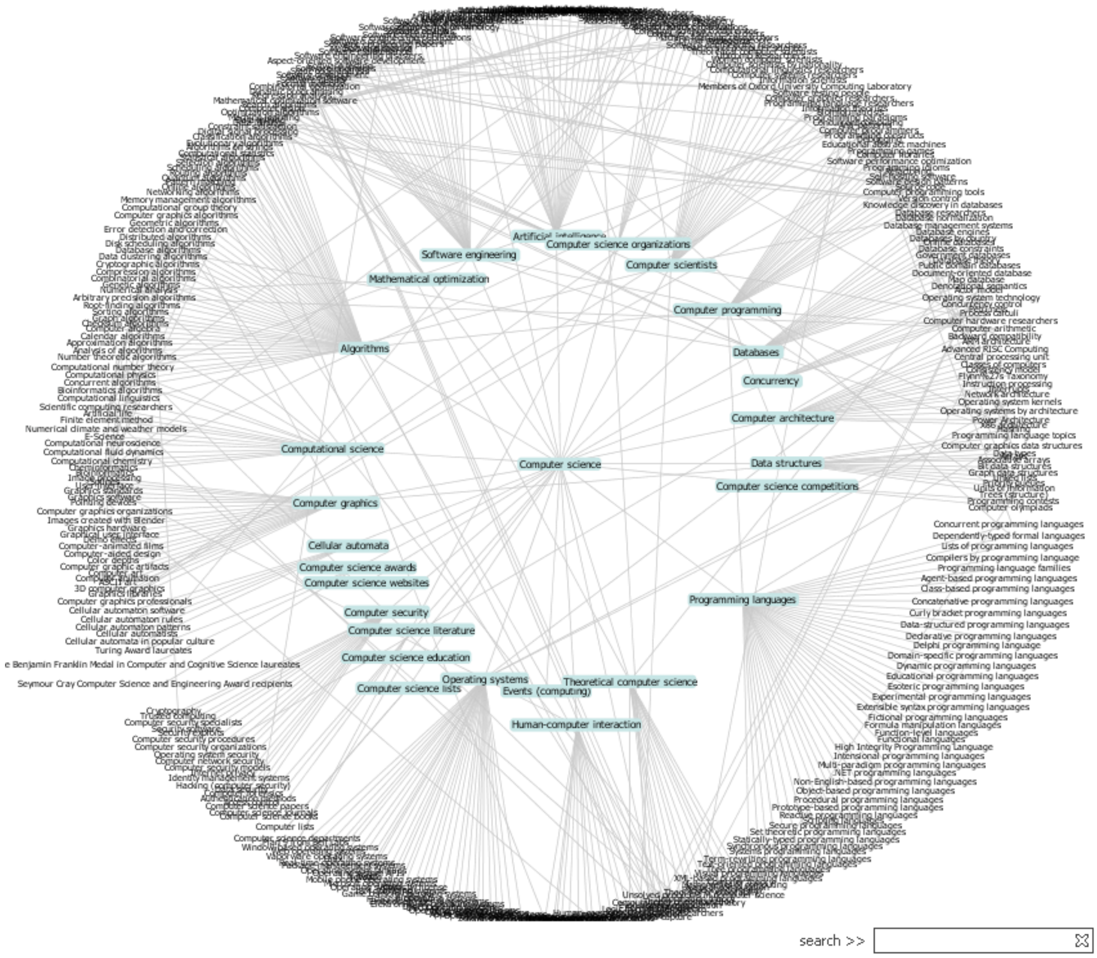

Figure 3.1.2-1 Visualization of the Categories Structure (Computer Science)

[4] A complete list can be found at http://www2.hawaii.edu/~pcchang/ICS699/results.html or http://www.cdpa.nsysu.edu.tw/~pica/results.html

Figure 3.1.2-23.1.2-3 is a closer look at certain categories using SpaceTree (Plaisant, Grosjean, & Bederson, 2002) to emphasize the hierarchal relation for more than 2 levels of depth. For example the highlighted branch has the following hierarchal relation: computer science > computer engineering > computer architecture > computer arithmetic > data unit.

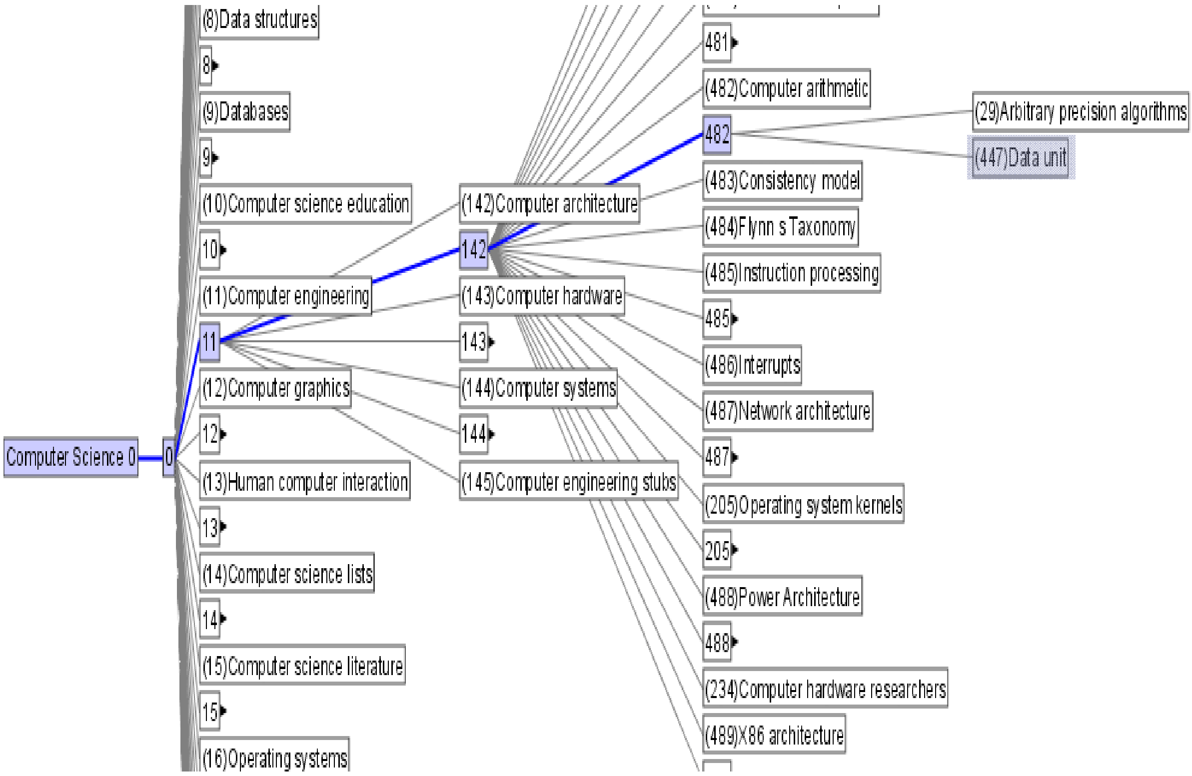

Figure 3.1.2-23.1.2-3 Visualization of the Computer Science Category Hierarchy

Heuristics for Keyword Extraction (Phase One)

There are two phases regarding augmenting the categorical structure with keywords: 1) heuristic keyword extraction and 2) keyword weighting. In the first phase of keyword extraction, the categorical structure is used as an ontological space to be augmented with keywords extracted based on content semantics and structure. We identify pages belonging to the CS category and its subcategories (378 in total) and define certain heuristics, listed below, to extract keywords from these pages.

1. Anchor texts, or structural terms (referred to as AT thereafter) from the content of each page – An indicator of the “is_a”, “has”, “related”, or “subclass” relations.
2. Representational terms or marked terms (referred to as RT thereafter, i.e., bold and italic terms) of each page – An indication of emphases made by authors; may capture the semantic information of a page.
3. Page title, or semantic terms (referred to as TI thereafter) – Capture the main topic of a page.
4. High TF-IDF terms, or statistical terms (referred to as TF thereafter) – the well-known VSM statistics on page abstracts – A brief summary with a few paragraphs and hyperlinks in XML format.
   (Page abstracts are available from Wikipedia’s data dump[5].)

Due to high frequency, some administrative texts in Wikipedia, such as “history”, “applications”, “fields”, and “see also”, etc. were ignored as part of a stop-words list to avoid any bias.

Figure 3.1.2-4 is a Wikipedia page about software engineering. Certain examples of each type of heuristic keyword are annotated in the page.

[5] http://download.wikimedia.org/

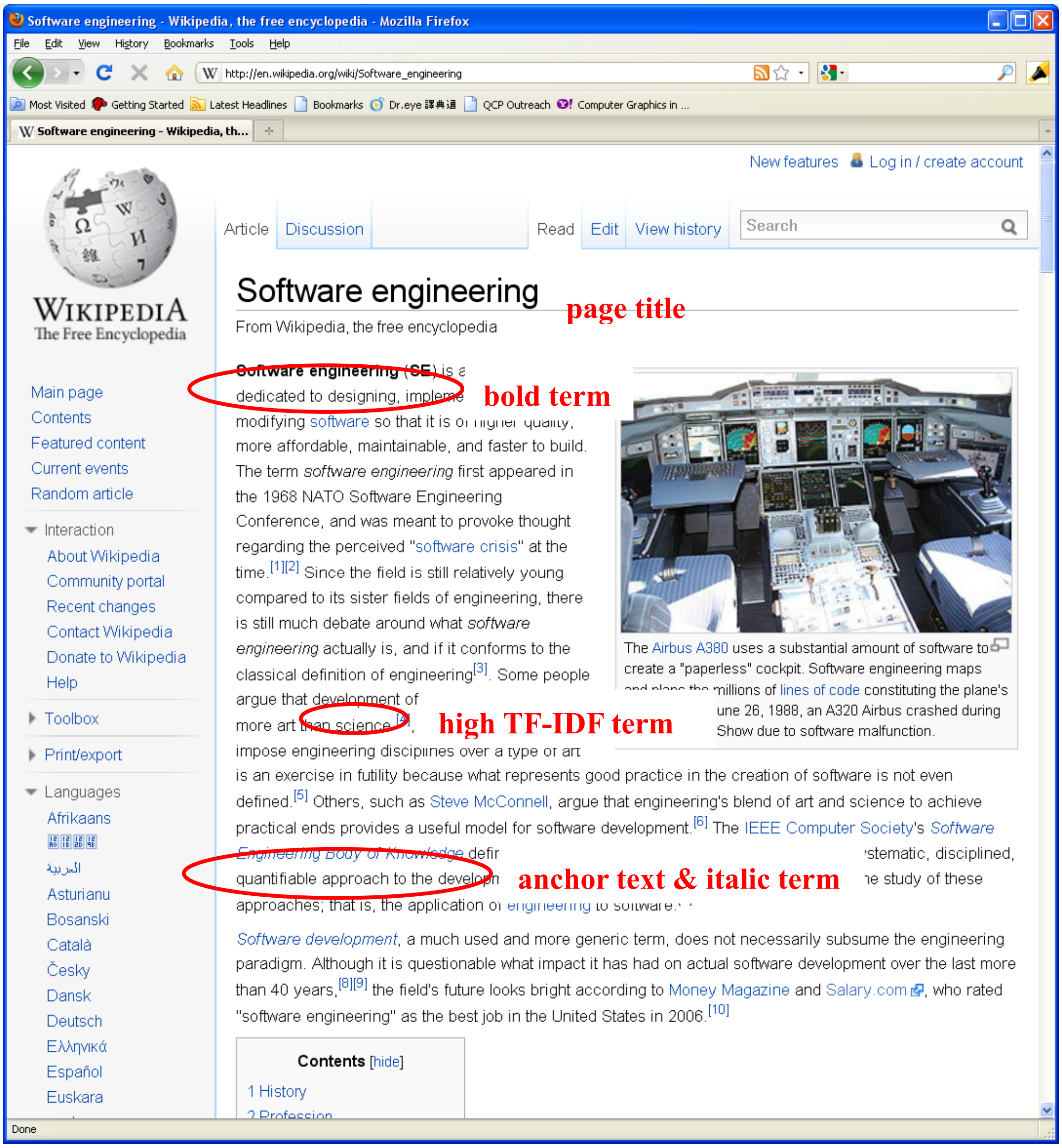

Figure 3.1.2-4 A Sample Wikipedia Page

In general, we targeted and extracted the above four types of heuristics as contextual keywords in the category to which their pages belong. About 50 keywords were selected to attach to each category. The augmentation is straightforward. For example, the bold terms extracted from a page under the “Algorithm” category were inserted into the “Algorithm” ontological category as keywords. In other words, we used the categorical relation as a vertical frame in the ontological space initially and then expanded the space horizontally by inserting AT, RT, TI, and TF keywords subsequently. Therefore, there are four ways to augment the space. The final WO is a combination (unification) of all of these heuristic keywords with weighting. The next phase explains this process.

Keyword Weighting (Phase Two)

In the second phase of the keyword augmentation, the expansion is a comparison, combination, and optimization of the above four types of heuristics. Keywords are weighted differently to reflect the importance of the common keywords that appear in multiple heuristics. Additionally, keyword weighting allows for the preference of specific heuristics in order to obtain the optimal result for page modeling. Below are the formulas for keyword weighting:

Definitions:

$|K_i|$ = the count of unique keywords (K) extracted for heuristic type i.

i = 1 to 4, which corresponds to AT, RT, TI, and TF heuristics

$freq(K_{ij})$ is the frequency of the j^{th} keyword K_{ij} in the collection K_i of heuristic type i, $1 \le j \le |K_i|$

$max(K_i)$ is the maximum $freq(K_{ij})$ in the collection K_i of heuristic type i, $1 \le j \le |K_i|$

Normalizing keyword weight:

$W(K_{ij}) = 100 \times freq(K_{ij}) / max(K_i)$, $1 \le j \le |K_i|$ for i = 1 to 4 ………………….. (1)

Because the same keyword may appear in multiple heuristics, keyword weights are combined across AT, RT, TI, and TF heuristics. The weighting formula for a keyword K_n is as follows.

$W(K_n) = \sum a_i W(K_n)$ for $i = 1$ to 4, $1 \leq n \leq |\cup(K_i)|$, ………….……………. (2)

where a_i is an optimized coefficient assigned to heuristic type i based on experimental results, which is explained in the section 3.1.3 "Training the Sensor".

Table 3.1.2-1 displays a portion of the anchor text (heuristics #1) extracted from pages belonging to the computer science and algorithm categories. As the algorithm is a subcategory of computer science, there are some duplicate keywords (computer science, algorithms, etc.). The number in the front indicates a relative weighting W_{ij} within a category calculated using formula (1).

Table 3.1.2-1 An Example of Keywords Extracted from the Anchor Heuristic

Computer Science Category	Algorithm Category
100 computer science	100 algorithms
57 artificial intelligence	40 algorithm
36 computer programming	43 computer science
32 computer graphics	23 set
32 cryptography	12 doi
25 programming languages	6 donald knuth
25 software engineering	6 dynamic programming
25 alan turing	6 cryptography
21 algorithm	6 machine learning
21 algorithms	6 production
21 bioinformatics	6 randomness
21 cognitive science	6 recursion
…	…

For each of the 378 categories in WO, there is a list of keywords with keyword weights in WikiBase. Eventually, we combined these four heuristics and optimized the keyword weights, which are explained in the next section (3.1.3).

WikiBase stores WO as well as content and user models. In this section, we have just described WO, the most important data type. The second data type stored in WikiBase is the models of a user's logged behavior. It includes user-provided usage data, system-formulated user models, and the interaction data (e.g., recommendation or feedback) between the user and the iPK system. The sensor is the main component that manages these models.

3.1.3.Sensor

The sensor formulates a user model based on the categorical topics of usage pages. This user model is a vector of available categories in WO. We call it the Wikipedia Ontology Model (WOM). The content model of each usage page is also represented using WOM in WikiBase. WOM is similar to the vector space model but categorical labels are adopted instead of terms as vectors in the model. The length of the vector in WOM is the number of categories belonging to a selected topical domain (e.g., CS). This avoids the problem of huge term vector lengths considering the large number of terms possible in web pages.

Approaches for Managing Content and User Models

The sensor maps a usage page into the content model (WOM) by sensing any keywords that appear in the page based on each category's keyword list, and then associates the page with the corresponding categories according to each keyword's weight. Whenever the sensor maps a usage page into a content model, it increases the score of the page's relevant categories in the user model and normalizes the score by the number of usage pages. Therefore, if a user accesses a specific categorical topic multiple times or accesses multiple pages of a similar topic, the user model will score higher in the corresponding category. Thus, the user model constantly evolves on the basis of usage visits.

Mapping a Usage Page to WOM

The sensor's main task is to identify the degree of association, called an Association Score, between an unknown web page and a categorical domain (CS in our case), utilizing those extracted keyword lists in WO for string matching. The following describes the way iPK calculates the Association Score to one of WO's categories given a sample web page. The sensor calculates a page's Association Score R_{sc} of all categories in WO.

Definitions:

$|C|$ = the number of categories in WO

$mfreq(K_{sic})$ is the match frequency of the i-th keyword K_{sic} in category c that matches a term from sample s. $1 \leq i \leq |\cup(K_j)|$, for j = 1 to 4, $1 \leq c \leq |C|$

For example, during a sensing task, $mfreq(K_{sic}) = 1$ when K_{sic} is identified as a match in category c for the first time; $mfreq(K_{sic}) = 2$ when the same keyword K_{sic} is identified as another match in category c for the second time. The final value is normalized by dividing it by the number of words in a document.

For a sample web page s, its Association Score R_{sc} to a category c in WO is:

$R_{sc} = \sum d \times W \times (K_{sic}) \quad \alpha^{mfreq(K_{sic})}$,

$1 \leq i \leq |\bigcup (K_j)|$, for j = 1 to 4, $0 < \alpha < 1$, for c = 1 to |C| ………………………. (3)

Function W is defined in formula (2); α is a degradation rate that de-emphasizes the importance of repeated matches; d is a coefficient to discount scores for partially matched keywords in cases when keywords contain more than one term and only a single or a few terms are matched ($0 < d < 1$ for partially matched keywords, or $d = 1$ for fully matched keywords).

Training the Sensor

We applied Open Directory Project[6], referred to as ODP hereafter, to train our sensor for optimizing keyword weights. ODP is a website on which users create a web directory of arbitrary WWW pages. Every listed page in ODP has at least one category defined by its registered users. Due to the fact that we are only interested in the degree of association between a web page and a category in WO, ODP is sufficient to train our sensor regarding a soft classification of web pages. We use the "algorithm" category in WO to illustrate the sensor's training process, which is applied to 10 selected categories in WO that co-exist in ODP.

1. We randomly selected m URLs from the "algorithms" directory and its sub-directories in ODP as root pages.
2. For each root page, remove noisy data and obtain "informative blocks" (Debnath, Mitra, & Giles, 2005). Informative blocks are the thematic content of a web page that excludes administrative texts, navigation links, banners, and advertisements, etc.
3. Calculate the Association Score R_{sc} based on sensing informative blocks of each root page. The average R_{sc} among m pages then becomes a reference value.
4. Adjust the a_i coefficient in formula (2) and the d value in formula (3) to make sure R_{sc}, scores highest in the algorithm category for all m pages.

Repeat step 1 to 4 as another iteration. Continue the iterations until the average R_{sc} for each iteration becomes stable, i.e. the standard deviation is small. Then the stable value becomes a threshold to determine if a page is relevant to the

---

[6] http://dmoz.org/

(algorithm) category or not.

Regarding the adjustment of coefficient a_i in step 4, we tested it using only a single heuristic initially. We found that the anchor texts (AT) heuristic performed slightly better and the representational terms (RT) heuristic performed slightly worse in identifying the categorical topics of a page. Afterward, we switched to an equal value of a_i for each type of heuristic and found that this created a superior performance compared to applying any single heuristic alone. Based on the equal value of a_i, we finalized the optimization by increasing the value of a_i slightly for AT and decreasing the value of a_i slightly for RT.

The tuning process for a_i could be further improved using machine learning techniques if more overlapping categories from both ODP and Wikipedia are available. However, the major focus of this dissertation is to test the feasibility and the performance of our proposed approach, rather than identify the optimal system configuration of the process. Therefore, we will leave optimizing the keyword weighting as future work.

Even though only 10 categories were tested, results indicate that the threshold values all fall into a range within ±15% of the mean. We therefore used the minimal value within the range as the standardized threshold value for all categories due to the lack of co-existing categories in both WO and ODP. Once the threshold values for all categories are identified, the sensor uses them to reference the associations between a user’s usage data and WO.

The Sensor and WOM

An accessed webpage is processed first to remove noisy data as described in step 2 above. Then, the sensor matches keywords that appeared in informative blocks of the page with those heuristic keywords in WO. Given a cleaned page, the sensing process produces a list of the Association Scores corresponding to each category in WO, as the WOM. Categories with scores passing the threshold value are predicted as relevant categories.

Our sensor applies a novel way to categorize web pages and identify relevant categories. The process does not require the immersive participation of domain experts to label a training dataset, which is a common method. In our approach, those pre-labeled pages in ODP replace this process. In addition, it should be noted that since we have applied heuristics to select keywords from web pages, more heuristics, such as authoritative thesauruses can be added to our paradigm as needed.

3.1.4.Crawler

The crawler's main purpose is monitoring and discovering. Monitoring ensures the awareness of pages in the watch list (either user-specified or system-inferred) as well as the "freshness" of WikiBase to reflect category or page changes. Therefore, iPK's crawler routinely performs update checks and relevant discovery. The crawler adopts the breadth-first strategy based on available anchor URLs of an assigned root page, such as the page from the watch list. Auxiliary pages (i.e.: pages linkable from the root) are fetched as discovered web pages, which are processed by the sensor to formulate their corresponding content models to be stored in WikiBase.

3.1.5.Matcher

The matcher primarily provides a digest of web page recommendations. Two types of pages are listed separately as a single digest: new recommended pages (new recommendation, Type 1) and updated pages in a predefined scope (update notification, Type 2). The matcher fetches the content model stored in WikiBase of every discovered web page. It compares the content model with the user model based on cosine similarity. For a relevant match (i.e. a high cosine similarity score), the matcher en-queues it as a recommendation candidate for Type 1. Any update of pages in the watch list, reported by the crawler, also yields an en-queue for Type 2 recommendation. Additionally, the matcher manages the interaction between a recommendation and a user's response.

Learning in Recommendation: Diversity Index

Recommendations for Type 1 are primarily based on the content similarity with the user model. In addition, we define Diversity Index to represent the "interest coverage" dimension of a user's possible interest areas based on WO's structure and quantify it as a variable. Interest coverage captures the semantic association among various areas of interest to the user given a domain. This was inspired by one of our pilot experiment (Chang & Quiroga, 2010a) that studied and modeled user interests. The experiment revealed that factors for modeling performance could be related to the coverage of a user's interests. Therefore, we designed the index to influence the strategies of recommendations. We intend to distinguish a user with diverse interests (e.g., machine learning, databases, and network) from another user with more focused interests (e.g., machine learning, artificial intelligence and data mining). Users with diverse interests receive recommendations in more topics.

The importance of interest coverage is dual. On the one hand, the coverage of a user's interests may imply the user's expertise in the domain. Page recommendations may consider the expertise level of a page and avoid recommending an introductory page to an expert user even if the page is semantically relevant. On the other hand, we intend to increase the possibility of information encounter and discoveries (Erdelez, 1997) through our recommender system. An information encounter is broadly defined as the chance or memorable experience to gain valuable or interesting information out of a random exploration.

Calculating Diversity Index

Diversity Index is calculated by applying the Minimal Spanning Tree (MST) concept based on the ontological structure, or the topology, (e.g. A in Figure 3.1.5-1) of available categories in WO. Figure 3.1.5-1 is used to illustrate the calculation throughout this paragraph. Each category is a node in the tree. iPK identifies user interests among categories with scores higher than a threshold value as Identified Nodes (e.g. ML, DB, SE and DM). Other categories that link Identified Nodes together, such as a parent or child, are called as Connecting Nodes (e.g. CS). The distance between an Identified Node and an immediately Connecting Node is 2 (e.g. d (DB, CS) = 2) while the distance between two adjacent Identified Nodes is 1 (e.g. d (DB, DM) = 1). A MST is constructed by traversing *ALL* Identified Nodes of a user in WO. An MST example is the bolded edges and the Identified Nodes in Figure 3.1.5-1

(B or C). DI is calculated by summing up all distances of edges in the MST (e.g. 2+2+1+1=6). Additionally, the system locates clusters in the tree by grouping adjacent Identified Nodes together in order to understand the broader topics of the user's interests better (e.g. grouping ML, DB, and DM as a cluster DB' of size 3). Following the above procedure, when a user's Identified Nodes are updated due to new usage pages, iPK re-calculates the DI value accordingly.

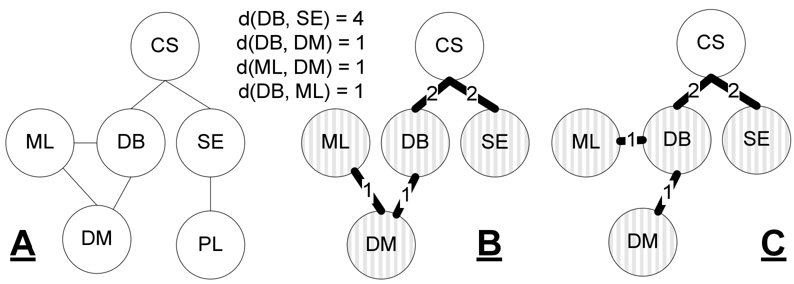

Figure 3.1.5-1 An Example for Calculating the Diversity Index

DI is used to prioritize recommended pages, which are selected by the Cosine Similarity measurement from pages fetched by iPK's crawler through usage crawling and querying search engine results. Using DI is our main strategy to diversify recommendations. The topic of a recommended page is alternated among all identified topics of potential interest to a user, according to the user's DI value. For example, if a user has a DI value of 90%, the system will have a 90% chance of switching the next recommended page from one topic to another within all identified topics. As another example, if a user's interests are more specific e.g. a DI value of 35%, the system will tend to recommend the same or similar identified topics. Therefore, a user with a higher DI value will receive recommendations in more topics. Considering DI in recommendation strategy, the system provides topical recommendations based on the interest coverage of a user.

In this chapter, we have introduced the iPK system and explained its major components – the knowledge base, sensor, crawler and matcher. The next chapter discusses the evaluation method and results.

Chapter 4 Methodology, Evaluation and Result Analysis

This chapter describes the methods, results and discussion of the evaluation of the iPK system. iPK has been implemented to model web pages, capture a user's topical interests through web pages, and match unknown web pages with a user model. The core system components were tested in three formative studies, followed by a pilot study. The modularized tests enabled the adjustments of individual components in order to improve the overall system performance. The main research question "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*" was then evaluated by 25 computer science (CS) working professionals. They provided rating judgments regarding iPK's recommendation's topicality and serendipity. Additionally, iPK's recommendations were compared with those recommendations generated by the pure content-based vector space model (VSM).

We expected similar performances from iPK and VSM regarding topicality, and a better performance from iPK than from VSM regarding serendipity. Additionally, we were interested in seeing how the Diversity Index (DI) measures an individual participant's topical coverage within the CS domain. Results indicate that iPK's performance is slightly better than VSM's regarding topicality, and significantly better regarding serendipity. A further investigation reveals that iPK is able to identify serendipitous recommendations that VSM may fail to recommend. iPK's superior performance in serendipity is possibly due to the augmentation of Wikipedia's categories with keywords, as well as the utilization of the categories' topology in DI. Furthermore, most of the DI values reflect a participant's coverage in the CS domain.

However, we were unable to identify any pattern of the DI values because of the small sample size.

Details of the evaluation method and result analysis are organized as follows. The variables for our main research question are operationalized in section 4.1. Evaluation of system components and the corresponding variables are described in section 4.2 (formative and pilot studies) and 4.3 (formal evaluation), followed by conclusion in section 4.4.

4.1. Operationalizing Research Questions

Figure 4-1 displays the visual relations between our research questions' independent and dependent variables. There are three levels of dependency: system, interaction and user. We examined the system-interaction and interaction-user levels of dependency sequentially in our formative studies.

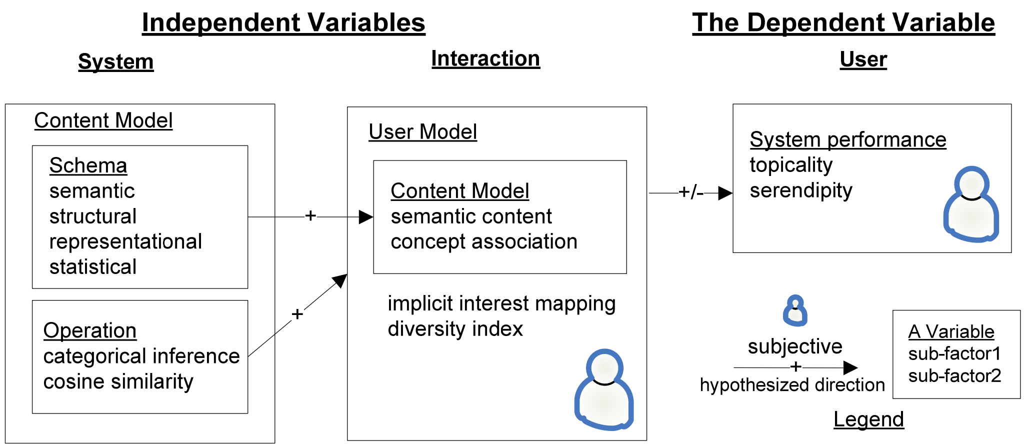

Figure 4-1 Visual Model of Research Variables

Operational definitions

A *content model* of a web page is defined as the data structure and values that reflect the associated Wikipedia categories of the page content. Schema and operation are two major factors for content models. A web page may be translated into different content models given different schemas. In the iPK system, we have applied heuristics to formulate schema. Configuration(s) of schematic parameters are optimized to map the topics in a web page into appropriate categories. The following list defines sub-factors that have implications for the performance of the content model.

- S*chema:* an interval variable of the keyword weighting method, which is operationalized as formula (2) in Chapter 3. Heuristics are defined in section 3.1.2.
- *Operation*: a nominal variable that is defined as a method to manipulate (create or compare) content models. *Categorical inference* is for mapping a web page to a schema and thus generating a content model for the page; *cosine similarity* is for a pair wise comparison between two pages' consent models.

User Model is divided into three sub-factors that influence the model performance:

- *Content model*: the WOM defined in section 3.1.3 as a list of categories.
- *Implicit interest mapping*: learning and accumulating the content model of usage pages with Association Scores described in section 3.1.3 as the main user model. Scores are based on an interval scale. The user model is aligned with the content model.
- *Diversity index*: a system-calculated interval variable for capturing the semantic association among various areas of interest to the user given a domain. The index is defined in section 3.1.5

System performance is measured through the following two perspectives:

- *Topicality*: measures the subjective judgment of how relevant a recommendation is to the user's topical interest. Topicality is an ordinal scale.
- *Serendipity*: measures the subjective judgment of how novel and interesting a recommendation is compared to the user's prior knowledge in his/her interest domains. Serendipity is an ordinal scale.

4.2. Formative and Pilot Studies

The purpose of the formative studies is to verify participants' viewpoints with the predictions generated by relevant system components and examine the flow and logic of the research design to evaluate the whole iPK system as well. Testing system components individually has the following advantage: "the effects of each component can be separated from one another and the weakest component(s) can be improved to create an effective adaptive system," as described by Chin and Crosby (2002, p. 108) The same participants were employed to test individual components and ensure the cohesiveness of the evaluation. Eventually, they evaluated recommendations generated by iPK after evaluating individual components.

Recall the main research question, QA "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*" We have investigated relevant subcomponents of the iPK system through three questions with the corresponding formative studies, which are explained in section 4.2.1 to 4.2.3. We then conducted another pilot study to test the flow, explore confounding variables, and estimate the quantity of usage pages, ratings, and the length of experiments for the formal evaluation. This is described in 4.2.4. The formal evaluation of the main question is discussed in section 4.3.

The three questions for formative studies are listed below. These studies were published in the indicated citation following each of the questions.

1. Can our content model correctly identify the topics of a web page (Chang & Quiroga, 2009)?
2. Does our content-based user model semantically capture a user's interests (Chang & Quiroga, 2010a)?
3. Does Diversity Index measure the coverage of a user's interests in the computer science domain (Chang & Quiroga, 2010b)?

The three questions sequentially explore the system's content model and user model on which the sensor and the matcher rely. On the one hand, the content model depends on Wikibase and the sensor so we only examined the content model, which is addressed in Q1. On the other hand, testing the user model in Q2 actually pre-screens the accuracy of iPK's recommendations by evaluating the predicted user interests. Q3 scrutinizes the core component of the matcher that accounts for serendipity in recommendations. The last pilot study tested the evaluation method regarding recommendations, which pre-tested QA.

The computer science (CS) domain was selected for all tests due to its relatively rich categories and pages in Wikipedia. We selected a seed category "computer science" and crawled out its subcategories up to depth 2, which produced 378 categories[7]. Regarding the sample population, we recruited 2 CS working professionals (at least 2

[7] A complete list can be found at http://www2.hawaii.edu/~pcchang/ICS699/results.html or http://www.cdpa.nsysu.edu.tw/~pica/results.html

years of experiences) with CS master degrees as our participants for all studies. They were asked to perform assigned tasks and answer survey questions. Details are described in section 4.2.1 to 4.2.4.

4.2.1. Content Model Generated by Sensor

Question 1: Can our content model correctly identify the topics of a web page?

Figure 4-2 displays the scope of variables examined in the red ellipse. Question 1 was addressed in this formative study. The schema for the tested content model is a weighted combination of each of the four heuristics.

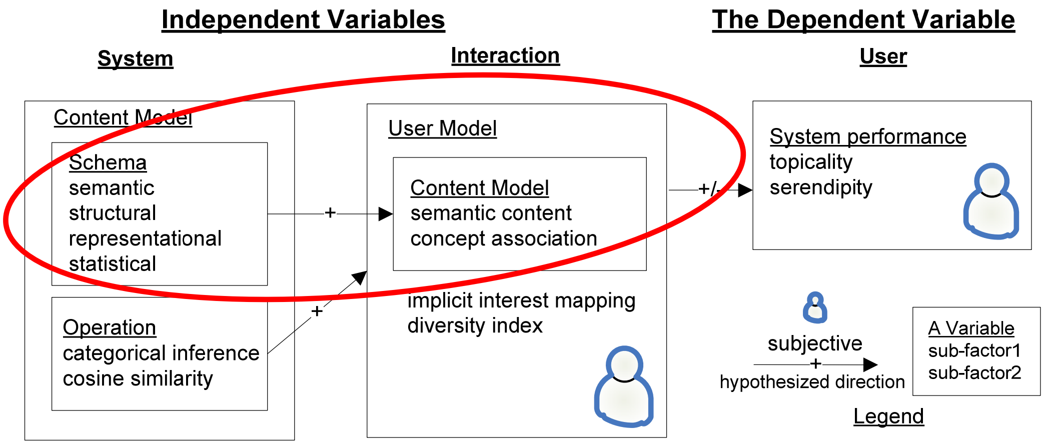

Figure 4-2 Scope of Sensor and Content Model Evaluation

We inspected the sensor's predictability in classifying web pages instead of testing WikiBase's structure directly. Two preliminary tests were conducted with the two CS professionals. In the first test, the following two pages were evaluated for their topical relevance. More pages were evaluated later in the second test.

A. http://www.algosort.com/, a directory page about computer programming

B. http://tc.eserver.org, a directory page about technical communicators

A list of classification results was shown to the participants and they had to order and rank the list. Considering the top two predicted categories, the system sensed and classified page A as “algorithms” and “genetic algorithms”; the system sensed page B as “human-computer interaction” and “usability” categories. In the evaluation of the classification result, both participants’ rankings orders are the same as the system’s ranking, for the top two.

In the second test, fourteen pages were selected, listed in Appendix A, from four topical areas – human computer interaction (HCI), data mining, algorithm, and computer games. These pages are directories, tutorials or organizational websites. Both participants had to evaluate five categorical keywords for each page. They evaluated the following statement "The given phrase is a topical keyword of the page" on a scale from 1 (strongly disagree) to 5 (strongly agree). The phrase is a categorical label generated by the sensor for each page. The following table summarizes the average ratings for each area.

Table 4-1 Ratings on Topical Keywords

	Participant 1	Participant 2
HCI (4 pages)	4.22	4.27
Data Mining (4 pages)	3.95	3.40
Algorithm (3 pages)	3.88	2.83
Games (3 pages)	2.67	2.2
Average	3.67	3.2

Discussion

From the result, the order of each area’s ratings from high to low is: HCI, Data

Mining, Algorithm, and Games for both participants. Excluding the Games topic, the average ratings are around 4 for participant 1 and 3.5 for participant 2. We suspect the wide topical or genre diversity of computer games could be the reason that the iPK system performed worse in that area. Another similar reason may be because the nature of computer games is tailored to various populations or domains, and thus requires different techniques or knowledge. Keywords predicted by the system reflect the common scientific techniques or theories (e.g. machine learning, or decision tree) for producing computer games, which are different from those tested pages that discuss computer games from a player's perspective. Other page modeling errors are possibly due to the fact that certain common keywords appear in many categories with a different weight, or maybe due to the frequency of matched keywords in web pages (too high). In summary, the above results indicate that our derived ontology may be a fair content model that maps unknown web pages to their related topical categories, especially for pages that do not have wide topical coverage.

4.2.2. User Model Generated by Sensor

Question 2: Does our content-based user model semantically capture a user's interests?

Figure 4-3 displays the scope of variables examined in the red ellipse. Question 2 was explored in this formative study. The schema for the tested content model is a weighted combination of each of the four heuristics.

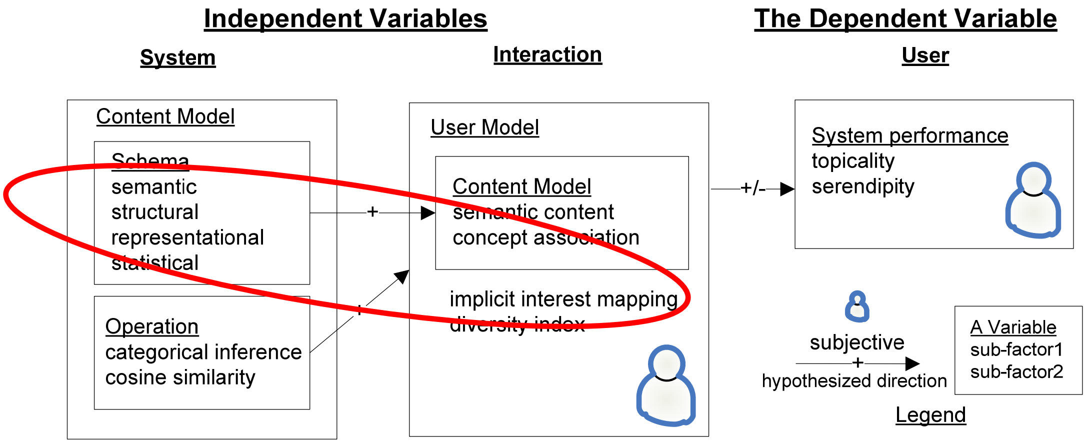

Figure 4-3 Scope of Sensor and User Model Evaluation

This formative study employed the same two CS professionals to evaluate the profile accuracy in capturing their interests. Both participants had to explicitly state their interests by selecting a list of 18 general categories in CS from Wikipedia. This was for evaluation purposes only and we did not want to overload the participants with too many categories. They also had to provide certain web pages (20) in CS that they visit frequently as the usage source to formulate the individual's user profile. After the automated analysis, individuals had to rate their predicted topical interests generated by the iPK system. We compared the explicitly stated interests with the predicted ones and analyzed participants' ratings.

The predicted interests for participants are divided into two sets – general and specific. iPK collected the general set by selecting categories from the list of 18 general categories that the participants used for stating their interests. The specific set includes any of the 378 categories from the WikiBase. To evaluate the specific set, we asked the participant to rate "Do you agree that you are interested in the following topic?" on a scale of 1 (strongly disagree) to 5 (strongly agree). Table 4-2 displays the results of the general set and Table 4-3 displays the specific set.

Table 4-2: Evaluation on General Interests

Explicit (#1)	Predicted(#1)
computer programming software engineering databases programming languages data structure algorithms computer security artificial intelligence	artificial intelligence software engineering human computer interaction concurrency theoretical computer science computer programming operation systems computer architecture databases data structure
Explicit (#2)	Predicted (#2)
artificial intelligence computer programming data structure databases human computer interaction programming languages software engineering	databases operation systems data structure artificial intelligence theoretical computer science computer programming concurrency computer architecture human computer interaction computer security

Table 4-3 Evaluation on Specific Interests

Participant # 1's Average Ratings: 3.52	
Image processing 2	Databases 5
Theoretical computer science 2	
Database management systems 5	
Computer programming 5	Security exploits 3
Machine learning 4	
Declarative programming languages 4	
Computer graphics 2	Dynamic programming 4
Human-computer interaction 3	Concurrency 4
Data structures 5	Events(computing) 1
Operating systems 3	Regression analysis 3
Document-oriented databases 4	Algorithms 4
Optimization algorithms 3	Instruction processing 3
Central processing unit 2	Access control 5
Programming constructs 5	Database theory 4
Artificial intelligence 3	
Participant # 2's Average Ratings: 3.92	
Databases 5	Checksum algorithms 2
Declarative programming languages 4	
Database management systems 5	
Systems programming languages 4	
Identity management systems 5	Image processing 4
Algol programming language family 2	
Curly bracket programming languages 2	
Software performance optimization 5	
Computer systems researchers 4	
Object-oriented programming languages 5	
Package management systems 3	Functional languages 4
Procedural programming languages 5	
.NET programming languages 5	
Extensible syntax programming languages 4	
Esoteric programming languages 3	Operating systems 2
Text-oriented programming languages 5	Software metrics 5
C programming language family 4	Pattern matching 5
Free software operating systems 3	
Multi-paradigm programming languages 3	

Discussion

Table 4-2 presents the explicitly-stated and the system-predicted interests for participants #1 and #2. We did not restrict the number of categories that the participants could choose but the system only predicts the top 10 categories. Answering question 2, Table 4-2 reveals that the system fairly predicts both users' general interests with minor errors. The prediction covers almost all of their interests. Therefore, based on the result from the two participants, the model seems to captures a user's semantic interests generally.

Table 4-3 shows more specific categories predicted by iPK and their ratings from both participants. The number following each categorical name is the participant's rating. The average rating is 3.52 for participant # 1 and 3.92 for participant # 2. As we can see, Table 4-3 more clearly identifies participant 1's interests in databases and programming areas and participant 2's interests in software and programming languages areas. This may reveal that the degree of interest varies among all explicitly selected categories. Although there are certain categories that seem misleading or missing in Table 4-3, the overall prediction performance is still acceptable with both average ratings exceeding 3.5.

An interesting phenomenon is that certain categories, such as "algol programming language family," are too specific for the participant's knowledge, particularly if the prediction is irrelevant. However, in relation to the concept of information encountering (Erdelez, 1997), a participant was surprised that certain predictions contain relevant terminology that was not associated by him before, e.g., associating “identity management systems” with “software agents.” Both concepts are of interest to the participant. It seems that the specific predictions can be extremely relevant or irrelevant. Nevertheless, results from both general and specific predictions indicate a need to examine the coverage of a user’s interest and the degree of concentration among interests, as they may be factors for recommendations.

We noticed that if the usage page is a portal or resource page that contains mostly URLs or directories, the iPK system tends to produce erroneous predictions. This may be a sign that the features, functions, or purposes, of a web page may influence the modeling performance. This issue is revisited again in the discussion of section 4.2.3. In addition, we also examine whether more usage pages yield better results to cover some missing categories and correct the errors in the process. In summary, the above results indicate that the content-based user model fairly captures a user’s topical interests with minor errors.

4.2.3. Diversity Index and Matcher

Question 3: Does Diversity Index measure the coverage of a user’s interests in the computer science domain?

Figure 4-4 displays the scope of variables examined in the red ellipse, which corresponds to question 3.

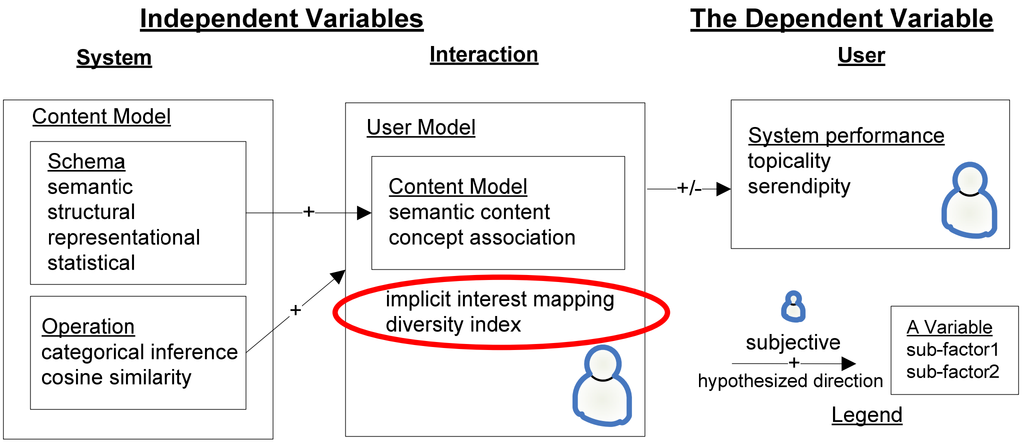

Figure 4-4 Scope of Matcher and the Diversity Index Evaluation

In this formative study, we were particularly interested in verifying the approach of using the WikiBase's ontological structure to study the coverage of a user's interests, which is measured with the Diversity Index (DI) on a scale of 1 to 100. DI was inspired by our previously mentioned experiment (Chang & Quiroga, 2010a) that studied and modeled user interests. The experiment revealed that factors for modeling performance could be related to interest coverage, as we distinguished general and specific user interests. Therefore, the iPK system applies DI as a factor to influence recommendations. Interest coverage is quantified as a variable by the ontological structure we derived, as was explained in section 3.1.5.

We created 8 simulated user models by respectively selecting 10 pages that represent each user's interests in the computer science (CS) domain. The complete list of pages can be found in Appendix B.[8] These 10 pages are either categorized in Open Directory[9] or are the top-ranked search results from Google, given the topic listed in

[8] Also available at http://www2.hawaii.edu/~pcchang/ICS699/umap.html
[9] http://www.dmoz.org/

Table 4-4. There are two kinds of topics: specific or hybrid. *We expect lower DI values for specific topics and higher DI values for hybrid topics.* In addition, we had two CS participants from the previous two formative studies, who provided 20 usage pages for user modeling and interest capture. Table 4-4 displays the DI values and the clusters with cluster size underlined. Larger cluster sizes imply stronger concentration.

Table 4-4 Diversity and Interest Clusters

#	Topics	DI	Interest Clusters
1	Database(DB)	41	DB9, CS organization3
2	HCI	35	HCI5, Graphics3, SE3, Computer architecture(CA)3
3	Algorithm	35	Algorithm13, Theoretical CS3
4	Software(SE)	77	SE5, DB3, HCI3, CA3
5	Data Mining(DM)	65	Algorithm8, Database4, Artificial intelligence(AI)3
6	SE\HCI	70	SE5, HCI4, DB3, CA3
7	SE\Algorithm	64	Algorithm9, SE4, Theoretical CS3
8	SE\DM	67	SE4, DB4, CA3, AI3, Algorithm3
9	Participant 1	65	DB6, Computer security4, Computer programming3
10	Participant 2	84	DB6

Discussion

Table 4-4 is discussed in three parts: specific topics (# 1-5), hybrid topics (# 6-8) and participants (# 9-10). We address certain features, such as functions, purposes, or information contained, of selected pages in the discussion. First, the simulated DB, human computer interaction (HCI) and algorithm "user" have lower Diversity values of 41, 35 and 35 respectively. The selection of DB includes pages from groups or companies that focus on database applications or systems. Therefore, the iPK system identified the CS organization cluster of size 3. HCI pages cover a variety of topics including graphics, software engineering (SE) and computer architecture (CA). Most algorithm pages are about algorithms introduction and usages so they are highly related to theoretical CS. The software and data mining "user" have higher Diversity values of 77 and 65 respectively. By nature, software and DM are application dependent and our page selection ranges from vendor pages, and explanations to application usages mostly for laymen, particularly the software pages. Therefore, this may be the reason that higher DI values are shown despite the fact that these pages focus only on a single topic.

Second, simulated users with hybrid topics have similar DI values to DM and software. This similarity indicates the mixed nature of different topics. Nevertheless, the number of clusters or the size of the cluster is generally smaller than a specific topic alone, unless there is an overlapping topic in the hybrid.

Third, both participants' DI values imply that they may be interested in more than two topics, particularly for participant 2 with a value of 84. However, the participant has a stronger interest in DB alone. Our earlier experiment recorded participant 1's explicit interests as: computer programming (CP), SE, DB, programming languages (PL), data structures (DS), algorithms, computer security, and AI. We further personally validated his focus in software programming and DB. Previous results also recorded participant 2's explicit interests in AI, CP, DS, DB, HCI, PL, and SE. In this study, participant 2 further revealed his focuses on DB and DM.

In general, the Diversity Index approximately reveals not only the coverage of a user's topical interests, but also identifies topical clusters that the user focuses on, both of which help comprehend the thematic interests of users. Answering question 3, DI approximately measures the coverage of a user's interest in CS. *Nevertheless, the index is an indicator but not an absolute value.* Potential sources of errors may occur during the page modeling process possibly due to the functional features of web pages. We found iPK tends to perform better for tutorial or introductory pages. These types of pages are closer to the nature of Wikipedia. We suspect that it may be beneficial for the iPK system to identify the function of a page before modeling, which is potential future work.

Additionally, our user modeling approach considers keywords and topic frequencies so the number of usage pages may be a factor as well. We suspect more usage pages may yield better accuracy. However, the time dimension is another concern as more usage pages are processed, since user interests may change and the degree of changes may vary according to each individual. All in all, we intend to emphasize the importance of interest coverage through quantification. Nevertheless, user's topical interests are not as simple as one variable since other factors are involved.

4.2.4. Piloting Recommendation Evaluation

We have discussed three formative studies in preparation for our main research question QA "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*" We conducted the pilot study to test the formal evaluation method for QA regarding evaluating recommendations generated by the matcher. The purpose of this pilot study is to test the flow, explore confounding variables, and estimate the quantity of usage pages, ratings, and the length of experiments for the formal evaluation. Figure 4-5 displays the scope of the system's evaluation. In this section, we revisit concerns of Q1 and Q2 explored in the first two formative studies as well.

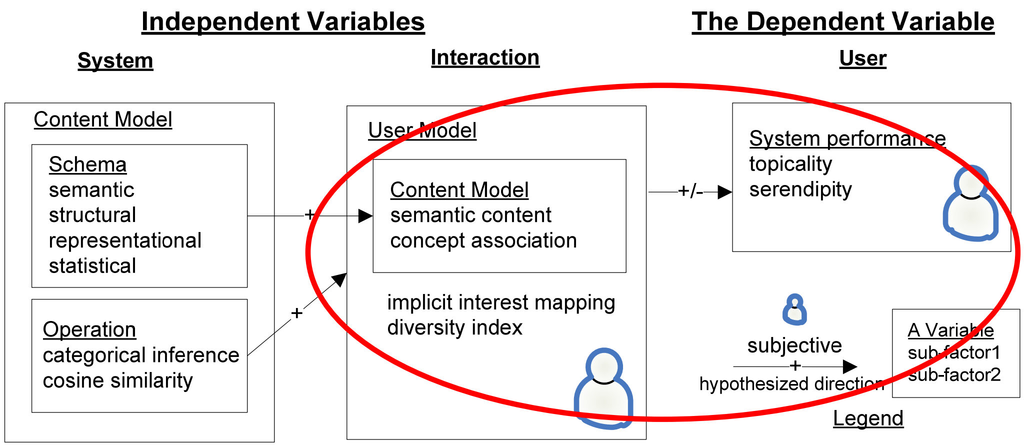

Figure 4-5 Scope of System Evaluation

Measurement

Recommendations were assessed based on user ratings with respect to topicality and serendipity. Topicality assesses whether a recommendation is related to the subject area of a user's interests (Xu & Chen, 2006). The question statement for participants is "I feel that this page is relevant to at least one topic of my interests – providing information related to the subject area(s) of my interests." Regarding serendipity, the earlier discussion in section 1.2 (problem statement) indicates that there seems to be no golden standard for measuring serendipity yet. Therefore, this dissertation borrows the definition of novelty, which is close to serendipity and is widely used in information retrieval. Novelty refers to the degree to which the recommendation is new to a user and beyond what the user already knows (Xu & Chen, 2006). The question statement for participants is “I feel that this page is novel -- providing new information to me and the information is beyond what I already know.” We define topically relevant and serendipitous recommendations as pages with high ratings in both topicality and novelty. Precision is used to measure relevance on topicality and novelty.

$$\text{precision}_{\text{Topicality}} = \frac{\{\text{relevant pages regarding topicality}\} \cap \{\text{recommended pages}\}}{\{\text{recommended pages}\}}$$

$$\text{precision}_{\text{Novelty}} = \frac{\{\text{relevant pages regarding novelty}\} \cap \{\text{recommended pages}\}}{\{\text{recommended pages}\}}$$

Participants

Regarding the experiment’s participants, we re-recruited the two from our earlier studies. We selected certain topics (Table 4-5) to cover some areas of interest to them. Both participants are capable of understanding all topics in Table 4-5 and are interested in at least one of them. This ensures that user ratings are based on the user’s areas of knowledge and personal interests instead of random evaluations. Participants provided their browsing histories (5 months) in advance with consistent visits to CS related pages from the logs of their preferred web browsers. The iPK system extracted CS related pages through the sensor component and then generated recommendations. Sufficient time is allocated for participants to read each recommended page before rating. To allow for potential rating explanations, participants rated recommendations by answering the survey questions in Appendix C using a think-aloud protocol. Questions include relevance ratings regarding topicality and novelty for each recommendation, evaluating categorical keywords of each recommendation, evaluating user interests predicted by iPK, providing the explanations for their ratings, and reporting their interests etc.

Table 4-5 Interest Topics

algorithms, artificial intelligence, databases, computer programming, informatics, computer graphics, programming languages, data mining, computer science award

Page Pool

Given usage pages, the system generated recommendations from a standardized page pool with the purpose of controlling the topicality variable and avoiding pages with complex or unfocused topics. There were 198 randomly selected web pages from various sources about Table 4-5's topics in the pool, with equal proportions of pages per topic. This selection included pages from search engine results and directory pages. We input the topics as query keywords in Google & Yahoo and obtained commonly appearing or top-ranked pages. Directory pages are classified as one of Table 4-5's topics in Yahoo Directory, Open Directory or Google Directory. Each participant was given at least 5 recommendations respectively from their past 1, 2, 3, 4 and 5 months of usage data. More recommendations were generated for longer usage periods based on the accumulated user modeled. This was intended to capture a wider range of possible topics through more recommendations, as longer periods may include changes in interest.

Table 4-6 Participants' Page Visit Frequency

Month	Visits (#1)	Accumulated	Visits (#2)	Accumulated
Feb. 2010	25	25	28	28
Jan. 2010	13	38	76	104
Dec. 2009	6	44	19	123
Nov. 2009	1	45	10	133
Oct. 2009	12	57	3	136

Results

Table 4-6 displays both participants' page visit frequency in the CS domain. Over the 5 months of usage data, there were 57 visits for participant 1 (referred to as #1 thereafter) and 136 for participant 2 (referred to as #2 thereafter). Table 4- 7 shows both participants' ratings on 35 (#1) and 40 (#2) recommendations respectively. Ratings are on a scale of 1 (strongly disagree) to 5 (strongly agree) for the Topicality and Novelty columns. iPK generated various numbers of recommendations, in column "# of Rec.", based on participants' past 1, 2, 3, 4 and 5 months of usage data. Taking ratings over 3 as relevant, the Precision of #1's ratings is 0.51 for topicality, 0.74 for novelty and 0.37 for both. #2's Precision is 0.8 for topicality, 0.35 for novelty and 0.31 for both.

Table 4-7 Relevance Ratings

	Participant 1		# of	Participant 2		# of
	Topicality	Novelty	Rec.	Topicality	Novelty	Rec.
Past 1 M.	4.00	3.80	5	4.00	3.00	5
Past 2 Ms.	3.67	3.50	6	4.17	3.60	6
Past 3 Ms.	3.63	3.63	8	4.14	3.60	7
Past 4 Ms.	3.63	3.63	8	4.11	3.60	9
Past 5 Ms.	3.63	3.63	8	4.08	3.60	13
Average	3.71	3.64	7	4.10	3.48	8
Sum	--	--	35	--	--	40

Precision (Topicality#1) = 0.51   Precision (Novelty#1) = 0.74

Precision (Topicality#1 ∩ Novelty#1) = 0.37

Precision (Topicality#2) = 0.8   Precision (Novelty#2) = 0.35

Precision (Topicality#2 ∩ Novelty#2) = 0.31

Table 4-8 Ratings on System-Predicted Page Topics

Participant	#1	#2
Past 1 M.	3.00	3.60
Past 2 Ms.	2.95	3.50
Past 3 Ms.	3.00	3.49
Past 4 Ms.	3.00	3.40
Past 5 Ms.	3.00	3.71
Average	2.99	3.54

For each recommended page, iPK generated 5 categorical keywords from the top ranked categories of its page model. Participants expressed their agreement on the keywords using the same scale from 1 to 5. This is the same task as what they did in the first formative study to evaluate Q1. Table 4-8 displays the average of ratings for each modeling period.

Table 4-9 Ratings on Top 5 User Interests

Participant	#1	#2
Past 1 M.	4.00	4.20
Past 2 Ms.	4.00	4.20
Past 3 Ms.	4.40	4.20
Past 4 Ms.	4.40	4.20
Past 5 Ms.	4.40	4.20
Average	4.24	4.20

For each modeling period, iPK predicted 5 topical interests for participants to evaluate the agreement on the same scale from 1 to 5. This is the same task as what they did in the second formative study to evaluate Q2. Table 4-9 shows the average of both participants' ratings.

Discussion

Topicality, Novelty, and Usage Data (Addressing QA)

Both participants' ratings for topicality are slightly better than ratings for novelty. Interestingly, #1 has a higher Precision value on novelty than on topicality. Conversely, #2 has a higher Precision value on topicality than on novelty. Judging from this result alone, we speculate that there could be a trade-off between topicality and novelty. This is a similar phenomenon to the trade-off of precision and recall, as described in section 2.2.6. Nevertheless, in the case of pages with low topicality and high novelty ratings, #1 said that “this is new and beyond what I already know, but I am not interested in it, ” when evaluating one recommendation. Although our definition of serendipitous recommendations requires high ratings in both topicality and novelty, we would consider modifying the question statement for question 2 to better measure serendipity alone. Furthermore, there is a need to compare iPK ’s ratings with another similar system’s to better evaluate the performance.

More usage data seem to increase the ratings on topicality. #1 had relatively more usage data in January compared to other months while #2 had more in February compared to other months. Therefore, #1’s average rating on topicality for recommendations based on the past 1 month is the highest (4). Likewise, the average rating on topicality for #2’s past 2 months is the highest (4.17). We previously suspected that more usage data would increase the accuracy of the system’s predictions. Even though the ratings based on monthly accumulated user models do not increase gradually, the amount of page visits is not proportional each month and user interests may change over the 5-month period. Therefore, given confounding variables, it is hard to judge whether more usage pages yield higher ratings on semantic relevance, particularly in topicality.

Days of Usage Data and User Interests Prediction (Addressing Q2)

One recent month’s usage pages seem to be sufficient for iPK to predict a user’s interests with a fair accuracy of 4.0 in topicality and at least 4.0 in system predicted interests, according to Table 4-7 and Table 4-9 (Past 1 M.). We informally tested predicted user interests based on daily, weekly and monthly usage data. We found that monthly data provided us the most converging topics in predictions, with slight to no changes in interest categories but gradual changes in categorical scores of the user model as more usage data were processed. This is reasonable because people may change their degree of topical interests gradually but topics that are of interest to them do not change on a monthly basis. According to the usage data, #1 is interested in databases and programming areas while #2 is interested in artificial intelligence, databases, programming, and human computer interaction areas over the 5 months. Interest predictions for both participants are verified as accurate as at least 4.0 out of 5, according to the ratings from Table 4-9. All self-reported interests are covered in the predictions for both participants.

Categorical Keywords (Addressing Q1)

There is no clear relation between the ratings on topicality and on categorical keywords of recommended pages, according to Table 4-7 and Table 4-8. The correlation coefficient of keyword ratings and topicality ratings is 0.25 ($p = 0.558$) for #1 and 0.2 ($p = 0.507$) for #2.

The page model is a vector containing the page's scores for every extracted category from Wikipedia in CS. Evaluated keywords are the top 5 categories that a recommended page belongs to. However, the differences of scores between the top 5 categories and other categories were not clearly reflected in the ranking. There may not be a significant scoring difference between the top 5 ranked categories and other categories. As a result, top-ranked categories on their own may not reveal all possible or representative semantic associations of a page and its related topics. We should examine the vector in its entirety and identify significant differences by semantically grouping categories together, according to Wikipedia's taxonomy of categories.

As we speculated earlier in Q1, the nature of topics (broad, specific, etc) may influence the content model's performance. Rating results of the categorical keywords were averaged to balance the potential effect. However, it would be more rigorous for knowledge experts to further define topical natures and examine the issue in detail, as potential future work.

Other Factors

Based on content analysis of the results from the think-aloud protocol, there are other factors that may influence the rating judgment. The participants mentioned the following factors: page length, timeliness (currency), the page author's perspective (opinion-related information), user availability (e.g. busy users tend to not like recommendations, which cause them information overload or "It is novel and interesting to me, but I have no time to learn"), and the focus of a page (e.g. pages about multiple concepts sometimes confuse readers).

In summary, the above results, with an average of 3.71 & 4.10 out of 5 in topicality ratings and 3.64 & 3.48 in novelty ratings, indicate both participants' positive reaction to the system's recommendations. Feedback from the pilot study helps us to make the following decisions for the formal evaluation. First, "interestingness" needs to be incorporated with novelty in the measurement for serendipity. We define it in the next section of formal evaluation. Second, we decided to contrast iPK's performance with the content-based VSM model that is widely used in information retrieval. Third, one recent month of usage data seems to be sufficient for iPK to model user interests with reasonable accuracy. We decided to track days of usage data and the frequency of page visits in the formal study, as they may be factors that influence the system's performance. Fourth, we decided to only use topical keywords to explore potential explanations for lower ratings. Fifth, confounding variables that may affect rating judgments need to be controlled as much as possible.

4.3. Formal Evaluation

QA. "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*"

We have piloted the main research question QA with an acceptable system performance and obtained feedback for designing the formal evaluation. Figure 4-6 displays the scope of variables examined in the red ellipse for QA. QA was evaluated using a weighted combination of the four heuristics for constructing the content-based model.

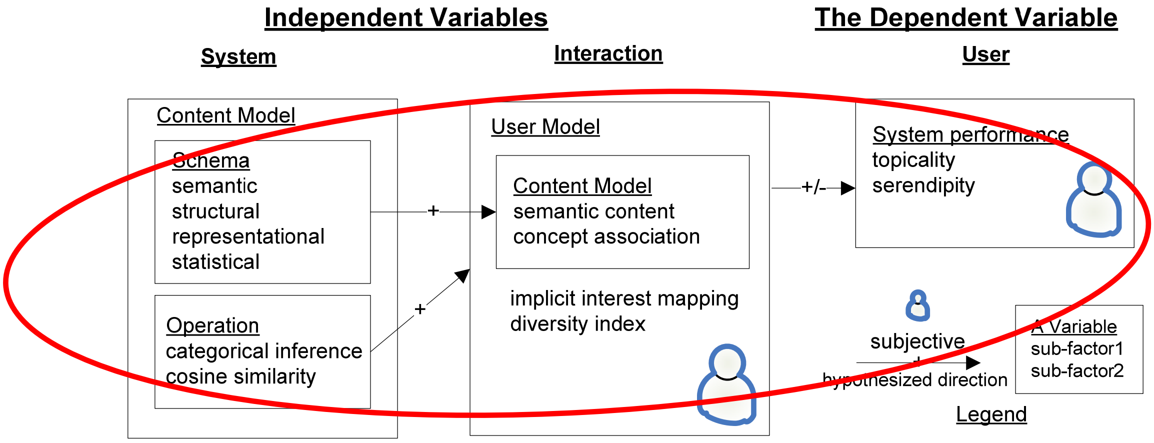

Figure 4-6 Scope of System Evaluation

Recommendations were assessed based on user ratings on a scale of 1 (strongly disagree) to 5 (strongly agree) with respect to topicality and serendipity. Topicality assesses whether a recommendation is related to the subject area of a user's interests (Xu & Chen, 2006). The question statement for participants is the same as the pilot study: "I feel that the information provided in this page is relevant to at least one topic of my interests – providing information related to the subject area(s) of my interests." Serendipity incorporates the concept of novelty and interestingness. Novelty refers to the degree to which the recommendation is new to a user and beyond what the user already knows (Xu & Chen, 2006). Serendipity infers novelty in a positive or interesting way. The question statement for participants is "I feel that the information in this page is novel – providing unexpected and interesting information to me." In this evaluation, we emphasize interestingness in the statement, taking into account the feedback from the pilot study. Precision, a standard metrics in information retrieval, is used to measure relevance on topicality and serendipity.

To contrast the difference of performance, we compared iPK's recommendations with the recommendations generated based on the classical vector space model (VSM) (Salton, Wong, & Yang, 1975) that applies cosine similarity as well.

Participants

We recruited 25 working CS professionals as participants to evaluate the main research question QA in an online survey. Participants were recruited through emailing college alumni of the computer science department at the University of Hawaii at Manoa. Most of them have a master degree, or have at least 3 years of working experiences as a programmer, a system designer or a network administrator. They were pre-screened and confirmed their interests in at least one topic from Table 4-10. Participants provided their browsing histories (mostly 1 month) from the logs of their preferred web browser. A log extraction tool was provided for them to select only CS relevant page visits as a list of URLs and timestamps. We examined systematic visits to CS related pages for all participants to avoid biased sampling, e.g. visiting numerous CS pages during only a short length of the whole usage period. We use the iPK system's crawler to fetch the extracted URLs for content modeling. User modeling requires the timestamp information.

Table 4-10 Interest Topics

algorithms, artificial intelligence, complexity theory, computer graphics, computational geometry, data format, data mining, databases, computer programming, computer science award, ecommerce, formal language theory, human computer interaction, informatics, information theory, information technology industry, management of information system, multimedia, mobile networking, parallel computing, programming languages, quantum computing, (computer/data) security, theoretical computer science, web hosting, and noisy topics within CS

Page Pool

Similar to the pilot study, the system generated 15 recommendations from a standardized page pool given usage pages with the purpose of controlling the topicality variable and avoiding pages with complex or unfocused topics. There were 632 randomly selected web pages concerning Table 4-10's topics in the pool, with equal proportions of pages per topic, plus pages in other arbitrary topics within the CS domain as noisy data. The selection process of the page pool is the same as the pilot study that relies on search engine results and directory pages. In order to control confounding variables, such as page length, or time-sensitive information, which may influence a user's ratings, we selected pages containing no more than 2000 words and had been updated within the past 3 years. Opinion-related or subjective pages were culled.

VSM's and iPK's Recommendations

In the VSM-based recommendations, we processed the one month usage pages as a single vector. All pages in the pool were converted into VSM as well. We selected recommendations by comparing the usage vector with every page vector in the pool and identifying the 15 most similar pages. Therefore, there were 15 pages selected respectively for the two types of recommendations – iPK system generated and VSM-based. Both types of recommendations were mixed randomly and presented together for user evaluation regarding relevance on topicality and serendipity. In the resulting analysis, we took the average of each participant's 15 ratings for both types.

To allow for potential explanations about ratings, participants provided a few topical keywords and optional comments. Additionally, they rated 10 categorical keywords predicted by iPK as their interests and reported a few keywords regarding their interests as well. These 10 ratings were averaged as well for each individual participant.

Our hypotheses include H1) there is no significant difference of the ratings regarding topicality between the iPK system and the VSM-based system, and H2) the ratings of the iPK system regarding serendipity are higher than the VSM-based ones. Regarding H1, we expected iPK to perform similar to VSM, and thus conceived the null hypothesis. As for H2, we expected a directional difference because iPK was designed to promote serendipity in recommendations.

4.3.1. Results

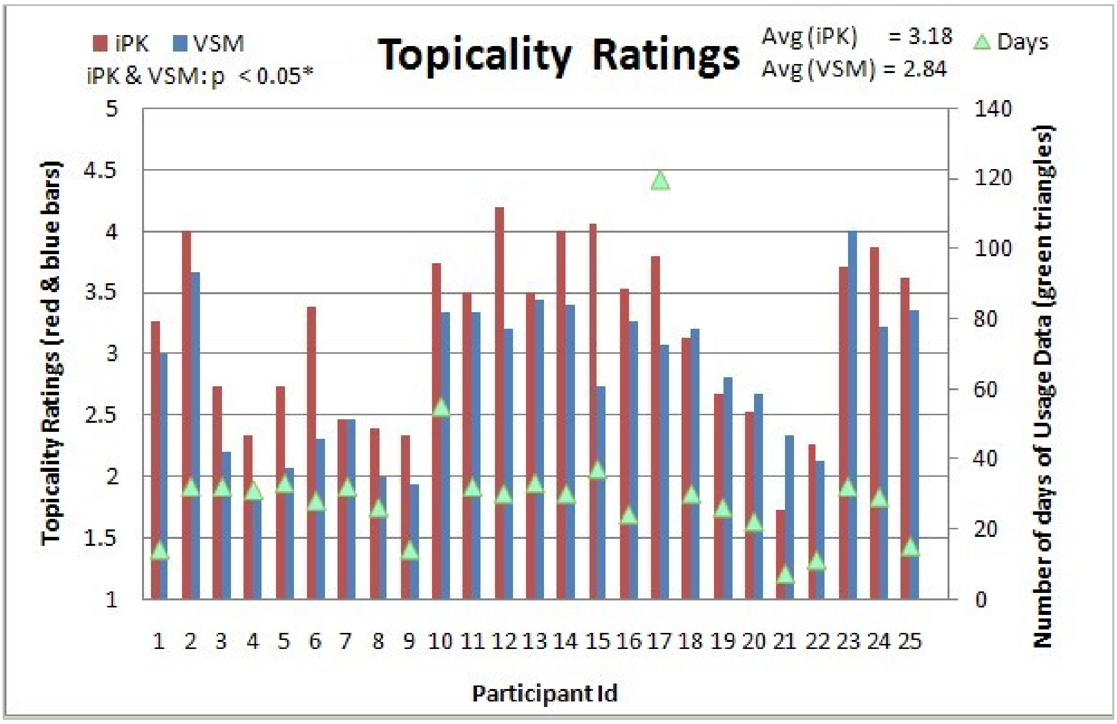

Figure 4-7 Ratings on Topicality

The average rating of the 25 participants regarding topicality is 3.18 for iPK and 2.84 for VSM. As for serendipity ratings, the average of the 25 participants is 3.11 for iPK and 2.80 for VSM. Therefore, iPK's ratings are higher than VSM's regarding both topicality and serendipity. We conducted both a paired t-test (T) and the Wilcoxon Signed-Rank test (W) to compare the rating difference between the iPK system and the

VSM-based system. The t-test requires a normal distribution while the Wilcoxon Signed-Rank test does not, both of which are commonly used in identifying the significant rating difference between two systems. Given the hypotheses of H1 and H2, we use 2 tail for testing topicality and 1 tail for testing serendipity. Both tests' results indicate a significant difference ($p < 0.05$) in iPK's performance in both topicality (blue and red bars using left axis in Figure 4-7, T: $p = 0.0007$, W: $p = 0.0013$) and serendipity (blue and red bars using left axis in Figure 4-8, T: $p = 0.0005$, W: $p = 0.0011$).

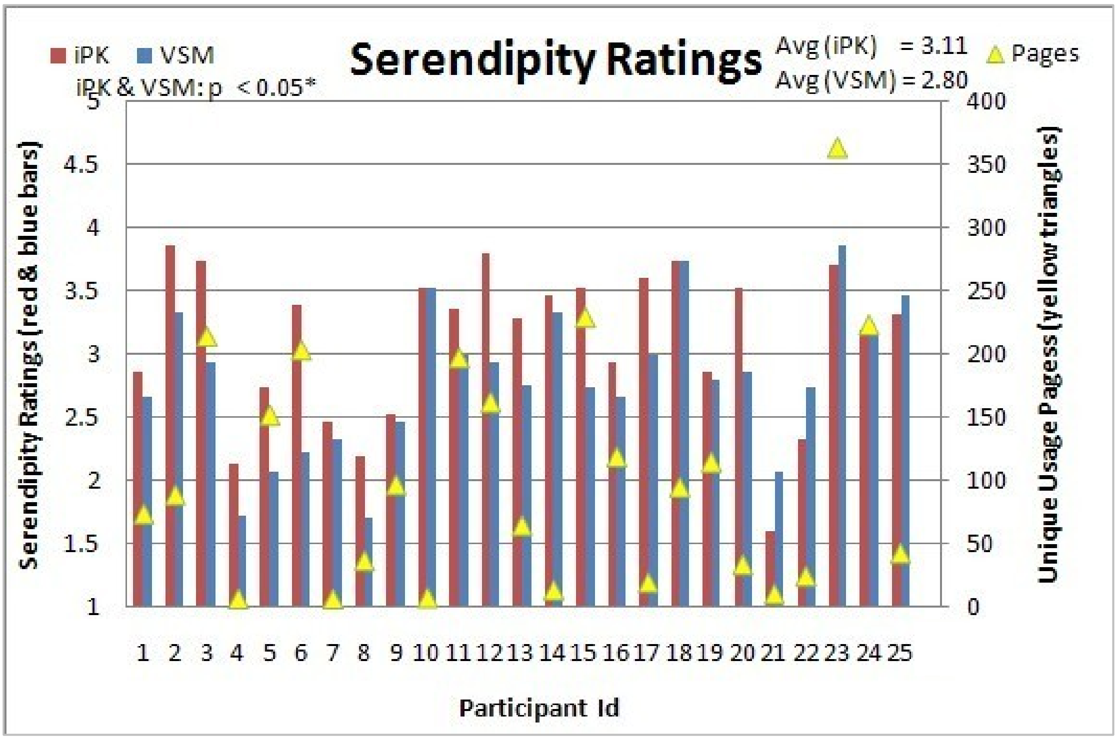

Figure 4-8 Ratings on Serendipity

Additionally, we examined potential factors that may influence the ratings on topicality and serendipity according to speculations from the formative and pilot studies: days of usage data (how many days), unique usage pages (ignoring revisits), and the amount of page visits (considering page revisits). Results show that there is a correlation between topicality ratings and days of usage data (R = 0.3942, p <= 0.0515) as well as another correlation between serendipity ratings and unique usage pages (R = 0.445, p <= 0.0262). In Figure 4-7, the green triangles using the right axis indicate the days of usage data for the each user. Similarly, Figure 4-8's yellow triangles using the right axis indicate the number of unique usage pages.

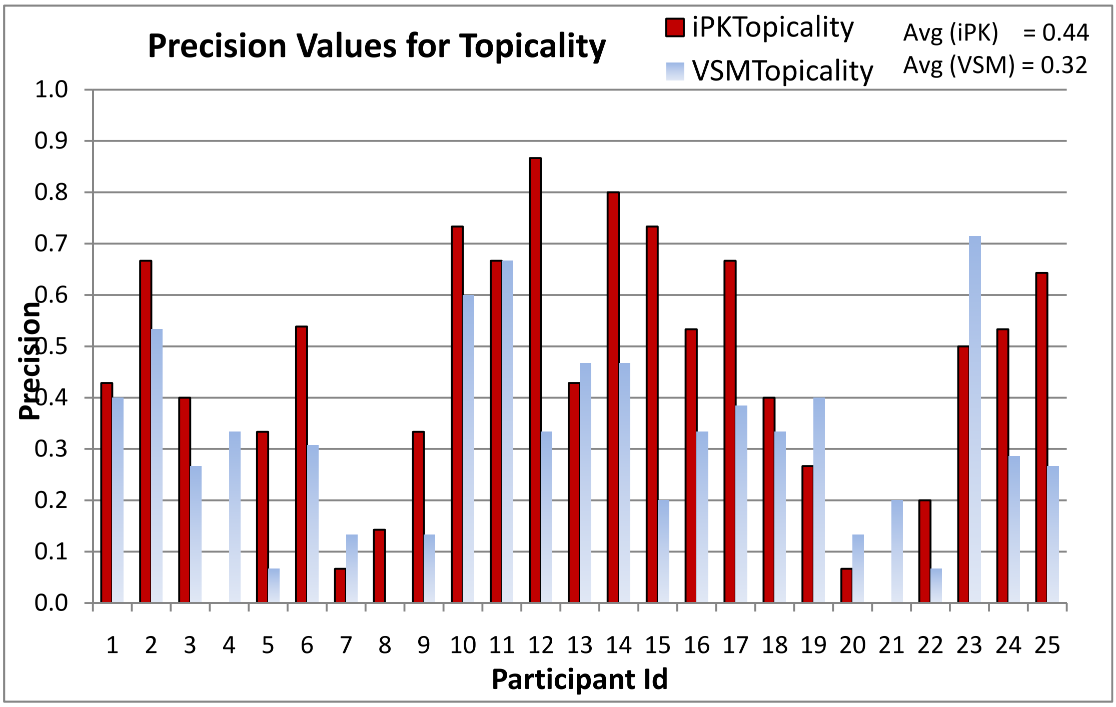

Figure 4-9 Precision Values for Topicality

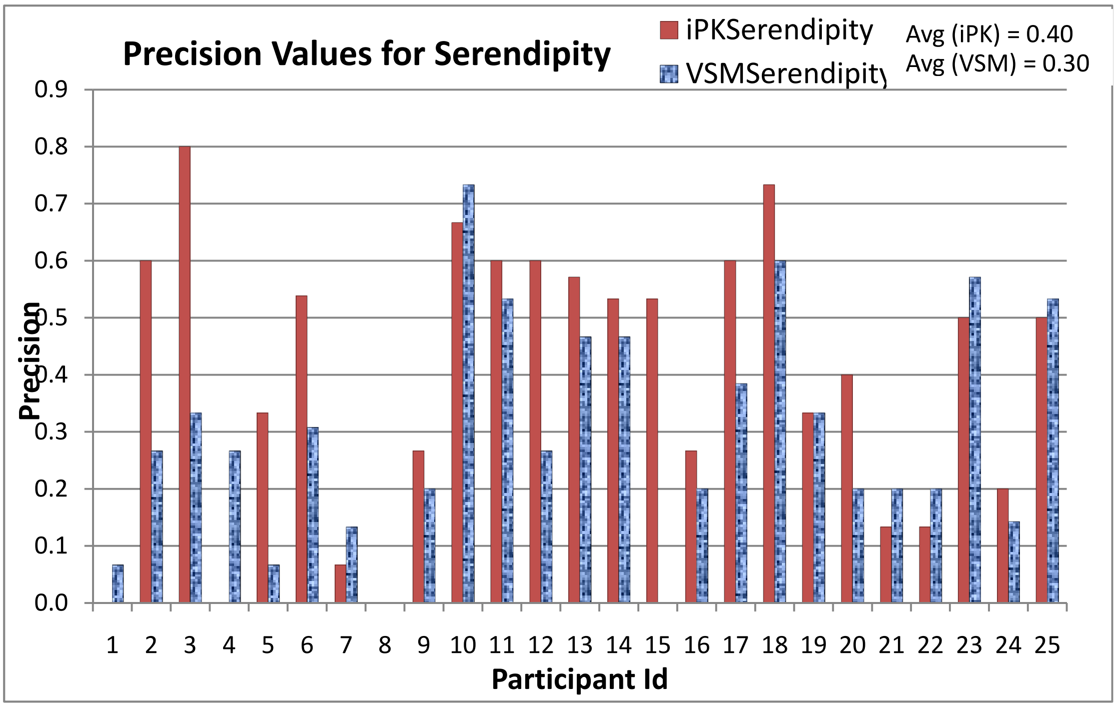

Figure 4-10 Precision Values for Serendipity

The precision values of individuals' ratings regarding topicality and serendipity of both systems are displayed in Figure 4-9 and Figure 4-10 respectively. Concerning topicality, the rating average is 0.44 for iPK and 0.32 for VSM. As for serendipity, the rating average is 0.40 for iPK and 0.30 for VSM. iPK's averages are higher than VSM using both measurements. In both figures, zero values are left empty in the field. Out of the 15 rated recommendations for both systems, we consider the ratings of at least 4 (agree) as relevant for the precision calculation. Again, both the paired t-test (T) and Wilcoxon test (W) indicate a significant precision difference between both systems for topicality (T: $p = 0.013$, W: $p = 0.017$) and serendipity (T: $p = 0.016$, W: $p = 0.030$).

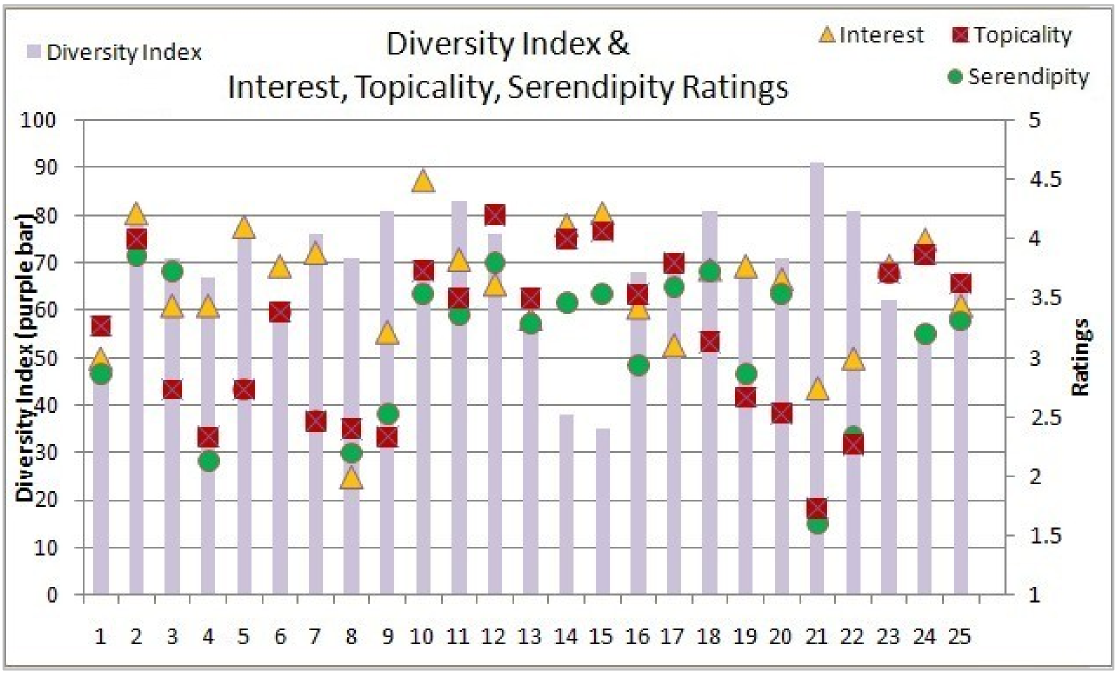

Figure 4-11 Diversity Index & Interest, Topicality, Serendipity Ratings

Figure 4-11 displays for each user the iPK system-calculated diversity index (DI) values using the left axis, and ratings on categorical keywords predicted by iPK as user interests, as well as ratings regarding topicality and serendipity on iPK using the right axis. The average of all 25 participants' DI values is 68.2. Regarding the categorical interest prediction, the average of all 25 participants' ratings is 3.58. Topicality and serendipity ratings are discussed in the previous paragraph but are listed again for reference purposes. We intended to investigate any interesting association among the four examined variables. Indeed, there is a correlation between topicality and interest ratings ($R = 0.5366$, $p <= 0.0059$). This is expected because recommendations are generated based on predicted user interests. Other than that, there is no clear relation among other combinations of the four variables. More analyses are explored in the discussion next.

4.3.2. Discussion

Hypotheses

H1 proposes that there will be no difference of the topicality ratings between the iPK system and the VSM-based system. Both the paired t-test and Wilcoxon Signed-Rank test (two-tail) indicate a significant difference among the ratings of the 25 participants. Therefore, H1 is rejected. Furthermore, H2 states that there is a directional difference of the serendipity ratings between the two systems. Applying the same tests (t-test and Wilcoxon Signed-Rank test with one tail) of H1 to serendipity ratings indicates a significant difference between the two systems. Therefore, H2 is accepted. Pictorially speaking according to Figure 4-7 and Figure 4-8, the iPK system generally performs better than the VSM-based system, except for a few participants. The next section explores the few cases.

Possible Explanations for Lower Ratings

Among the 25 participants, we explored those who rated iPK's recommendations lower than the VSM-based ones regarding topicality or serendipity. In Figure 4-7, participants 21 and 23 gave higher ratings to VSM-based recommendations than to iPK's. Similarly, participants 18-20 rated VSM-based recommendations slightly higher than iPK's regarding topicality. In Figure 4-8, Participants 21-23 and 25 rated VSM-based recommendations higher than iPK's regarding serendipity. According to both figures, participants 21 and 23 seem to prefer VSM over iPK. A closer look at their data, which are not precisely shown in both figures, reveals that participant 21 has only 11 page visits (all unique) over 7 days; participant 23 has 363 unique page visits (419 visits are considered revisits) over the 32 days of usage data, which is the highest number of visits. Both cases show the extreme sides of the usage sampling, which may account for the lower ratings.

Most participants provided approximately one month of usage data so the amount of usage data for participant 21 is obviously lower than others. Similarly, participants 22 and 25 only have 11 and 15 days of usage data respectively. *This may indicate that our system tends to perform worse in terms of topicality or serendipity in cases of too little or too much usage data.* Indeed, correlation analysis shows a relation between topicality ratings and days of usage data, as well as a relation between serendipity ratings and unique usage pages. Possible explanations are underfitting or overfitting of the usage data to the user model, which can cause lower ratings. Underfitting may happen when there are not enough representative data. Overfitting may happen when the model fits too much into the usage data that are related to short-term interests. Nevertheless, given the subjective ratings and limited sample size, it is difficult to generalize our speculation of the rating trend. Furthermore, the rating differences between participants using the same scale may cause certain biases, e.g. participant 2 tends to give ratings above 3 whereas participant 5 tends to give ratings below 3.

Additionally, usage data may include a few biased samples, as case 21 has shown. A further inquiry to participant 21 discloses that the usage data are not representative of his behavior. Data were not logged from his most preferred browser. Even though he provided one month of usage data, page visits that are relevant to the CS domain cover only 7 days. This is the danger of using self-provided data, but we compromised and accepted it due to privacy concerns. Participants may not trust installing "spyware" on their computers to monitor their website browsing usage, despite the fact that only CS relevant pages are our focus.

User Interests Prediction (Addressing Q2)

The iPK systems predicted 10 categories to represent a user's interests. The orange triangles in Figure 4-11 display the average ratings of the predictions for each participant. iPK seems to predict user interests well because most of the ratings are above 3 except for participant 8 (2) and participant 21 (2.75). Therefore, participant 8 has relative lower ratings on topicality and serendipity compare with other participants. The case of participant 21 was discussed earlier. However, a closer look at the ratings from participant 8 found an inconsistency in his ratings. His self-reported interests include "data mining" and "database structure," both of which were predicted but rated only as 3. A majority of participant 8's ratings are at most 3 out of the 1-5 scale for either interests or recommendations, with only 2 exceptions. As a result, removing the two outlier cases of participants 8 and 21, iPK generally predicts user interest with an average accuracy of 3.57 out of 5 for 23 participants. Nevertheless, the issue of rating inconsistency suggests future work to design system functions for inconsistency detection, explanation (e.g. interest changes), and resolution.

An interesting issue is that interest prediction lies in the specificity of WikiBase's ontological structure. Almost all general or broader topics of user interests could be correctly identified by iPK, but not those fine grained topics, because specific categories were omitted during our extraction from Wikipedia. To reduce the evaluation complexity, we selected categories-subcategories or categories–neighbors only up to depth 2. Therefore, WikiBase does not include fine grained categories for discerning specific user interests. This could also be due to the fact that the nature of taxonomy tends to group finer information together as a general category. Despite the lack of fine granularity, iPK alternatively identifies a broader category that fits better. For example, participant 11 is interested in "web application programming" while iPK identified "computer programming" and missed the "web application" scope. In view of this, metadata for categorical topics may be needed. Possible solutions include increasing the complexity of WikiBase's ontological structure, or analyzing the co-occurring topical interests and then inferring contextual metadata. These are potential future work.

Diversity Index (DI) and Interest Clusters (Addressing Q3)

Figure 4-11 does not disclose any clear relation between either topicality or serendipity ratings on iPK and participants' DI values. We did not have any hypothesis for participants' DI values and their ratings regarding topicality or serendipity due to the small sample size. Nevertheless, we were still interested in identifying any potential pattern from the data. All participants have a DI value raging from 50 to 91 out of 100, except for participants 14 (38) and 15 (35). Interestingly, participants 14 and 15 have relatively higher ratings on recommendations regarding topicality compared with other participants. Nevertheless, this phenomenon could be for the two cases only. On the other hand, participants 21 and 22 with a higher DI value of 91 and 81 respectively have relatively lower ratings on recommendations regarding both topicality and serendipity. Nevertheless, this is not the case for participant 11 who has a DI value of 81. Therefore, we could not conclude any potential relation, but only indentified a few interesting cases. Participants 21 and 22 are discussed again in a few paragraphs later.

Due to its length, the complete list of participants' DI values, clusters and reported interests is given in Appendix D. We only discuss certain interesting points here.

Generally speaking, most of the DI values and the clusters capture participants' coverage and interests. Certain participants' DI values may not truly reflect their interest coverage due to missing categories in WikiBase. This issue is revisited again in the next discussion regarding page topics. Take participant 23 for example. To our surprise, the 378 categories in WikiBase do not include "computer network." Therefore, participant 23 has a DI value of 62 even though he focuses only on the computer network area (expected to be less than 50 according to the formative study results in section 4.2.3). Nevertheless, rated categorical interests indicate his interests in other areas, e.g. HCI as well, but we suspect the degree of his interest in other areas is not as strong as in networking.

Participant 7 and 22 include clusters that seem erroneous, possibly due to the lack of representative usage data for their reported interests as we manually examined their 7 and 25 usage pages respectively. For participant 7, most pages used are from organizational websites; participant 22 browsed mainly technical news talking about different concepts. As a result, iPK predicted based on biased data.

There is another issue that is consistent with the formative study's results in section 4.2.3. Single topics that cover wider areas, e.g. computer games, tend to have higher DI values, despite the focus on a single topical area. For example, participant 3 seems to have a focus area of SQL databases within the context of business intelligence or related topics. However, business applications may be domain dependent to some degree and thus have a higher DI value of 71. Similarly, participant 21 has the highest DI value of 91 due to the broader topical interests in computer games and "surfing the Internet."

Generally speaking, identifying patterns among participants is difficult without a highly controlled evaluation procedure that recruits participants who are interested in similar topics and have similar topical coverage. Nevertheless, identifying a significant number of such participants is challenging due to their availability. Therefore, it will be left as future work.

Page Topics (Addressing Q1)

We studied the ratings regarding topicality and serendipity on iPK of all participants' 15 recommendations and identified the following points. First, for relevant pages that provide fundamental information of a topic (e.g. SQL standard specification for SQL), there is a tendency of lower ratings regarding serendipity among participants, despite the fact that the topic is highly relevant to them. This result is due to the same issue that iPK is currently unable to identify page topics to a finer degree. Therefore, identifying the level (e.g. beginning, intermediate, advanced etc) of information provided in a page is another area for potential topic of future work.

Second, participants' rating patterns vary. There are four participants whose ratings regarding topicality are at least 80% the same as ratings regarding serendipity. They understood the survey questions correctly and claimed that those topically relevant pages were also novel and interesting to them, if they had not seen those pages before. On the other hand, for pages discussing popular topics (e.g. cloud computing), certain ratings are low in topicality but high in serendipity. Some participants are not familiar with those "hot" topics; some others are interested in them only for future references. However, they all feel good to know those popular topics. This pattern occurs occasionally among participants.

Third, in a few cases, it is reasonable to speculate that the page pool does not provide any pages that cover a user's specific interests. For example, participant 9 reported her interests as "GIS, geo-visualization, and multivariate." The page pool does not include any of these topics. However, the closest topic related to visualization is computer graphics, which is indeed covered in her recommendations. Although not self-reported in the online survey, participant 9 was confirmed to be interested in computer programming during our recruitment prescreening. A similar situation related to missing topics happened to participant 21 when graphics related pages were recommended to him, as his interests of video games were not included in our topic pool. Ratings on these recommendations are relatively lower because of the "indirect relatedness." In view of this issue, a further measurement on recall is needed to reveal whether lower ratings are caused by missing topics in the page pool. However, it would be difficult for each participant to rate all 632 pages in the pool.

Recall Measurement

Recall measures the proportion of relevant recommendations out of all relevant pages in the pool. There are two types of relevance defined in our evaluation: topical relevance and serendipitous relevance. For the purpose of recall measurement, we selected a subset of pages in the pool that are potentially relevant to users. Participants 1 and 2 agreed to evaluate the subset of 140 and 120 pages respectively from the pool.

Excluding noisy topics, there are approximately 20 pages for each topic in Table 4-10. According to their interest indications, all pages from the identified topics of interest to them were evaluated. Again, we counted the ratings that are at least 4 (agree) out of 5 as relevant and considered the relevant recommendations among the 15 that had been evaluated earlier for both systems regarding topicality and serendipity. In other words, the recall value is the proportion of relevant recommendations out of the 15 recommendations over all relevant pages in the pool. Table 4-11 displays the two participants' recall values. Please keep in mind that only 15 pages were recommended and the possible maximum value for recall is not 1. For example, the maximum recall value is 0.375 (15/40) for participant 1 regarding serendipity.

Table 4-11 Recall Values for Participant 1 & 2

	Topicality			Serendipity		
Recall	System	VSM	Max	System	VSM	Max
Participant 1	0.167 (6/36)	0.14 (5/36)	0.417 (15/36)	0.275 (11/40)	0.225 (9/40)	0.375 (15/40)
Participant 2	0.159 (10/63)	0.127 (8/63)	0.238 (15/63)	0.667 (10/15)	0.267 (4/15)	1.00 (15/15)

iPK still performs better in the two cases for recall measurement, particularly in terms of serendipity. Nevertheless, recall may not be valued too much in recommendations considering serendipity. In fact, we intend to increase the diversity of recommendations by deliberately selecting topically similar pages with variances in topical coverage. Furthermore, it is possible for a topically irrelevant page to provide a user a sense of serendipity because unexpectedness (topical, in this case) is a part of the definition for serendipity. Another issue is that serendipity may depend on the sequence of presentation. A novel page among a list of topically similar pages may have relatively higher novelty than in a more diverse list of pages. To explore more about serendipity, another test was conducted to examine those serendipitous recommendations generated by iPK in the next discussion.

Examining Serendipitous Recommendations

Serendipitous recommendations refer to those pages that are rated at least 4 (agree) regarding serendipity. We intend to clarify the reasons contributing serendipity to iPK's recommendations. For each participant, we identified those serendipitous pages from iPK system that were not recommended by the VSM-based system and examined their cosine similarity scores. Results indicate that only a very small portion (0-33%, varying among participants) of these serendipitous pages have high similarity scores (within the top 50) for VSM. This suggests that the augmentation of Wikipedia's categories with page keywords and the algorithms are adding information from Wikipedia that simply cannot be captured purely by VSM. In other words, iPK seems capable of generating serendipitous recommendations that VSM may fail to identify.

Participants' Comments

Finally, we analyzed the participants' comments and identified the following points. First, serendipity is somehow related to the timeliness of a web page, especially for those time-sensitive websites, such as technical gadgets or news. Second, there is a perception difference of topical granularity among participants. Some participants define it as providing recommendations in very specific topics of their interests while others accept recommendations in general or marginal topics as relevant. This could be attributed to the reason that topical relevance is subjective and individuals' boundaries of relevance vary. Third, other page-related factors (e.g. timeliness, opinions, presentation style, authors/sources or context etc.) influence the ratings as well.

Optimizing Heuristics

iPK has been evaluated regarding its performance in prompting serendipity in recommendations. A further inquiry to another research question Qh investigated the optimized configuration of iPK's heuristics. Qh: "*Which means of constructing the content-based user model (semantic, structural, representational, or statistical heuristics) provides better recommendations that are relevant to a user's topical interests*" The usage data provided by the 25 participants were used to study the question. The following paragraphs report the findings of the investigation.

There is probably no difference among recommendations generated by these four heuristics. The precondition of Qh lies in the fact that different heuristics of constructing the content model provide different recommendations. However, we found that, among these four means, at least 47% of the recommendations are common for 80% of the 25 recruited participants in QA's evaluation. By these four heuristics, the system further predicted categorical keywords for each recommended pages individually. Similarly, we found that the results are just different orderings of the same keyword set for 84% of the participants.

For 7 participants, "representational" seems to have slightly different keywords but they tend to be erroneous, according to our judgment. *As a result of the reported numbers, assessing Qh is probably not necessary because these four means of model* *construction generate similar categorical keywords and recommendations.* We speculate that this is because the weighting formula in iPK does not considerably distinguish the four heuristics. Another reason could be that a significant proportion of keywords commonly appear in all four heuristics with just different weights. Nevertheless, this reason would only affect those web pages that contain the portion of keywords. Given the provided usage pages from each participant do not have equal amounts of these pages, no clear conclusion could be drawn even after closely examining the issue. Investigating the weighting formula and common keywords will be left as future work, as our research focus is to test the alignment of content modeling with user modeling using Wikipedia's content for semantic recommendations.

4.4. Conclusion

QA. "*Does our recommender based on Wikipedia's content provide topically relevant recommendations, promoting serendipity, of pages from different websites in a selected domain?*"

In conclusion and answering QA, iPK's performance on topicality and serendipity is significant better than the classical VSM model. This is possibly due to the augmentation of the ontological modeling representation in WikiBase, and due to the utilization of the topology of Wikipedia's categories to examine the coverage of user's interests. Recommendations from different websites are diversified and the performance regarding serendipity is thus better than the VSM-based system.

For the purpose of semantic user modeling, we emphasize the alignment of content models with user models using Wikipedia's content to achieve semantic relevance for recommendations. We have implemented a system to demonstrate our idea and test the hypothesis. The performance of iPK regarding topicality is slightly better than the classical VSM-based recommendations, and is slightly superior in terms of serendipity (both statistically significant). Therefore, we conclude that the alignment of content models with user models, through usage analysis, carries the semantic representation for topically relevant recommendations in the iPK system. In addition, our proposed approach relies on usage mining at the client side. This reduces user effort in formulating the user model and grants them more control.

Summarizing interesting points from the results and discussion, seven issues are listed below.

1. Capturing users' topical interests using keywords is easier than recommending pages of their interests, which involve confounding variables and subjective factors.
2. The amount of usage data influences the quality of iPK's recommendations.
3. The topology of Wikipedia's categories and its augmentation can potentially reveal the coverage of a user's interests.
4. Results from examining serendipitous recommendations suggest that the WikiBase we derived adds information from Wikipedia that simply cannot be captured purely by looking at word similarity with pages from usage data alone.
5. A finer granularity of more categories may be needed in WikiBase to capture more comprehensive areas or topics of web pages and user interests.

6. Types or features of pages, such as tutorial pages or blog pages, may influence the system's predictions, as explored in section 4.2.2

Certain other points are worthy of emphasis. First, using Wikipedia's categories as a vector to model a user's topical interests yields a simple and potentially interoperable model. This is an important concern in cross-system recommendations. Second, combining categories with heuristic information extraction for keywords leaves room for the selection of heuristics. In this study, we have demonstrated that the use of heuristics in Wikipedia's content is successful to some degree for semantic user modeling. Different domains or user groups are able to apply heuristics of their interest. Using the semantics encoded in Wikipedia is worthwhile as well. Third, we incorporated content-based modeling with usage analysis to minimize user intervention, which is required for other user modeling approaches, such as collaborative filtering. We emphasize the convergence between content modeling and user modeling as the formulation of a user-derived ontology. Fourth, our recommendation engine works on the client side, which eases the privacy concern of disclosing sensitive information to web service providers. Fifth, using categories and keywords to create an ontology is our attempt at content and user modeling, but by no means a replacement for an authoritative ontology like SUMO (IEEE, 2000) . Instead, Wikipedia is an alternative to complement the weaknesses of expert-constructed ontologies, such as their higher cost to construct and update. Additionally, user modeling and interest identification are our foci, rather than semantic accuracy of the ontology. We supply a methodology for researchers to further develop similar personalized agents.

Finally, while there are similar web page recommender systems, such as WebWatcher (Joachims, Freitag, & Mitchell, 1997), Syskill & Webert (Pazzani, Muramatsu, & Billsus, 1996), WebMate (Chen & Sycara, 1998), and Ultra Gleeper (Richardson, 2005), there are slight differences from the iPK system in terms of functional purposes. These systems actively provide a filtering mechanism in cases, such as information retrieval based on keyword queries, and web browsing assistance by predicting preferred links. This is different from our purposes of passively recommending pages by monitoring web usage data. Furthermore, they are all prototypes and slightly outdated at this time. Therefore, it is hard to compare iPK with them. We thus chose only the VSM-based system for evaluation.

Chapter 5 Future Work and Applications

The research platform constructed by this work is rich with opportunities for improvement. Suggestions for future work discussed in this chapter include: 1) the data structure, the maintenance, and the visualization of the knowledge base; 2) other factors that may influence the recommendations and ratings; 3) the representativeness of CS to other knowledge domains; and 4) content quality. We explore these more in the following paragraphs. Lastly, we relate our work with other applications as a closing remark.

1. **The Knowledge Base (WikiBase):**

As identified earlier in our evaluation results, increasing the granularity of the WikiBase ontology to include finer and related topics is potential future work. In addition, the ontological structure of the knowledge base provides associations to neighboring categories. It also enables the aggregation of categories at different levels of granularities using clustering algorithms such as K-means based on the neighboring relations. We can further make use of the linkage relation of Wikipedia's pages to enlarge categories that do not contain enough pages or information.

Investigating the keyword weighting and common keywords in WikiBase is another area worth investigation to obtain the optimal configuration, as explored in section 4.3. Moreover, the rules or mechanisms for identifying web pages' functions or purposes could be contextual metadata for WikiBase. This is because the evaluation results reveal a need to consider the level (e.g. beginning or advanced) of information provided in web pages for recommendations.

Furthermore, the maintenance of the knowledge base needs to be automated as efficiently as possible. The "freshness" of WikiBase relies on the crawler to check Wikipedia regularly for any modification. Although employing speedier machines can reduce the computing time, effective scheduling algorithms may need to balance the server workload.

Lastly, the visualization of WikiBase merits future work in order to explore "interest tracing" questions such as the following: How do we trace and visualize the changes in a user's interests given a certain time frame? How do we identify and visualize a user's persistent interests as opposed to short-term interests? Are there any relationship between a user's interest changes and the categorical topology, e.g., shifting interests to neighboring categories of current interests? Will frequency of recommendations influence interest change? How should the system respond to interest changes? These questions are more complicated due to the additional time dimension. Visualization may help us to comprehend users' interest changes.

2. **Other Factors that Influence Relevance Judgment:**

There are other user-related, document-related, and situational factors that may affect the recommendation judgment. Some of these user-related factors (Quiroga & Mostafa, 2002) are education, experience, working status, intention, lifestyle, emotions, beliefs, and privacy concerns; document-related factors are accessibility, language style, verbosity, currency, reliability, authorship, and quality of source; situational factors (Petrelli, Angeli, & Convertino, 1999) are time available and motivation. There may be certain conflicts or inconsistencies when considering all of these in user's judgment. Although this dissertation only focuses on semantic relevance, the above factors more or less influence the degree of interest in recommendations. A personal ontology (Quiroga, 2009) taking some of these factors into consideration may be of help in improving different aspects of the recommendations .

3. **Representativeness of the Computer Science Domain:**

In this work, we employed CS professionals to test our proposed methods. The representativeness of CS to other knowledge domains may be an issue. In this dissertation, we assume that CS professionals are familiar with the Internet and Web 2.0 techniques. They may tend to use the web to facilitate their careers as well as to satisfy their interests. Therefore, their usage data may more reflect their interests. Additionally, there are relatively rich categories and pages in the CS domain from Wikipedia. Therefore, whether can we apply the assumption and our proposed paradigm to other knowledge domains is still a challenge. Web page modeling may not accurately reflect those people who are not tech-savvy or those who use the Internet for limited purposes, e.g., email-checking. In addition, experts in other knowledge domains than CS may not contribute as frequently as CS professionals to Wikipedia due to, e.g. the lack of computer skills for editing Wikipedia.

Another related issue is the inequality of contributions to Wikipedia that was mentioned in section 2.4. Ortega, Gonzalez-Barahona, and Robles (2008, p. 6) studied Wikipedia and found that “a small percentage of the total number of authors (less than 10% of the total number of authors in all the editions we considered) are responsible for the majority of the total number of contributions to the encyclopedia.” Similarly, Priedhorsky found that the top 10% of editors, by number of edits, contributed 86% of the persistent word view, the number of times any given word introduced by an edit is viewed (Priedhorsky et al., 2007). This problem may occur in any domain.

An interesting area is to study the influence of content statistics, such as author numbers, page contributions, volume or the granularity of the categories, on iPK's performance. Not every domain in Wikipedia contains rich categories and articles like computer science. Therefore, the quality of recommendations may be related to some of the content statistics. Although comparing the proportion of contribution among different knowledge domains is not our primary focus, this issue certainly warrants investigation. There are also many other factors affecting the representativeness of the computer science domain.

4. **Content Quality:**

One last issue regarding Wikipedia is its content's quality. Wikipedia's content and categorization system play an important role in our method to generate recommendations. This work emphasizes the framework to automate the ontology generation and its performance in recommendations. Nevertheless, the quality of Wikipedia's content is still controversial. It will be worthwhile to adopt the same framework using domain experts to create Wikipedia's content and ensure the quality.

Finally, implementing the proposed approach fully as an application may give us more diverse and real world usage data other than the data provided by the CS professionals in the evaluation. In this case, any WWW page could be a potential recommendation, rather than the restriction of the "page pool." The WikiBase may be updated according to user behavior or feedback. We look forward to a large scale implementation and evaluation.

Closing Remarks:

This work focuses on personalization. Relating our work to others, we use a figure from the literature (Razmerita, 2003) to express the possible applications of our work. Figure 5-1 points out four directions of the roles of user models and user modeling in knowledge management systems, which also apply to our context. While we work only on personalization, there are still three other self-explanatory directions. The scope covers the interaction between (personalization) systems and systems, users and users, and users and systems.

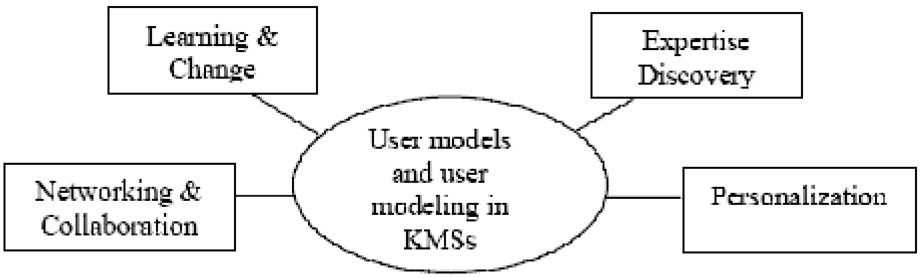

Figure 5-1 The Role of User Models and User Modeling in Knowledge Management Systems

To name some possible applications: first, related to learning users and detecting information changes, our agent system can integrate with another personal information management (PIM) system like "Stuff I've seen" (Dumais et al., 2003) and help users to locate a previously-seen web page. Any change in the page can also be identified.

Second, related to PIM, it is possible to construct a personalized information repository on the client side by making use of the personal crawler in the iPK system.

Users can decide what kinds of information will be crawled from various websites and displayed in their personal views.

Third, related to guiding users' learning, the iPK system could be a web browser plug-in that facilitates surfing the Internet for users. Visualization techniques like the floating window in the TechCrunch[10] website can be combined to facilitate the information foraging experience.

Fourth, related to networking & collaboration, our user model is basically a categorical keyword list functioning as a partial personal ontology. It can be readily transferred to another platform for system networking to formulate a more comprehensive personal ontology. More personalization services could be provided to users with their personal ontologies.

Fifth and related to expertise discovery, social networking with peer website users or service users is a popular topic despite privacy concerns. Privacy concerns could be reduced by granting user control to define the appropriate disclosure of their personal ontologies. Finally, other system applications or possibilities are left to the imagination of our readers.

[10] http://www.techcrunch.com/

Appendixes

Appendix A：Sample Pages for Question 1

Due to limited space, only the first page of each selected topic displays the categorical keywords.

Algorithms

http://www.algosort.com/

(Algorithms, Genetic algorithms, Root-finding algorithms, Networking algorithms, Disk scheduling algorithms)

http://www.oopweb.com/Algorithms/Files/Algorithms.html

http://cgm.cs.mcgill.ca/~godfried/teaching/algorithms-web.html

Data Mining

http://www.data-mining-guide.net/

(Databases, Algorithms, Knowledge representation, Natural language processing, Knowledge discovery in databases, Machine learning, Data mining)

http://www.thearling.com/

http://databases.about.com/od/datamining/Data_Mining_and_Data_Warehousing.htm

http://www.ccsu.edu/datamining/resources.html

HCI

http://www.pcd-innovations.com/

(Human-computer interaction, Human-computer interaction researchers, Usability, Computer science organizations, Artificial intelligence, Software development)

http://www.nathan.com/

http://nooface.net/

http://www.hcibib.org/

Games

http://www.robinlionheart.com/gamedev/genres.xhtml

(Image processing, Computer programming, Demo effects, Regression analysis, Computer graphics)

http://open-site.org/Games/Video_Games/

http://www.literature-study-online.com/essays/alice_video.html

Appendix B : Page List for Topics in Question 3

Topic(s)	URLs
Database	http://www.sqlrecipes.com/ http://www.vldb.org/ http://conceptoriented.org/ http://www.paragoncorporation.com/ArticleDetail.aspx?ArticleID=20 http://www.daniel-lemire.com/OLAP/index.html http://myweb.brooklyn.liu.edu/gnarra/database/ http://liinwww.ira.uka.de/bibliography/Database/index.html http://www.cs.wisc.edu/dbworld/ http://www.dbdebunk.com/index.html http://www.sigmod.org/
Data mining	http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm http://databases.about.com/od/datamining/Data_Mining_and_Data_Warehousing.htm http://www.tdan.com/view-articles/5263/ http://www.thearling.com/text/dmwhite/dmwhite.htm http://www.autonlab.org/tutorials/ http://www.statsoft.com/textbook/data-mining-techniques/ http://databases.about.com/od/datamining/a/datamining.htm http://www-users.cs.umn.edu/~kumar/dmbook/index.php http://www.data-mining-guide.net/ http://www.ccsu.edu/datamining/resources.html
Algorithm	http://www.algosort.com/ http://www.personal.kent.edu/~rmuhamma/Algorithms/algorithm.html http://www.nist.gov/dads/ http://editor.altervista.org/menu.html http://www.oopweb.com/Algorithms/Files/Algorithms.html http://www.softpanorama.org/Algorithms/index.shtml http://cgm.cs.mcgill.ca/~godfried/teaching/algorithms-web.html http://www.cs.sunysb.edu/~skiena/214/lectures/ http://www.personal.kent.edu/~rmuhamma/Algorithms/algorithm.html http://java.sun.com/docs/books/tutorial/collections/algorithms/
Software	http://www.analysttool.com/

	http://www.microtool.de/e_index.htm http://www.offshoresoftwaresuccess.com/ http://www.reqbib.com/adavis/ http://www.pera.net/Ind_tool.html http://www.bredemeyer.com/ http://www.construx.com/Page.aspx?nid=14 http://www.se-radio.net/ http://www-01.ibm.com/software/rational/ http://www.sei.cmu.edu/research/
Software \ HCI	http://www.pcd-innovations.com http://tc.eserver.org http://www.nathan.com http://nooface.net http://www.hcibib.org http://www.reqbib.com/adavis/ http://www.pera.net/Ind_tool.html http://www.bredemeyer.com/ http://www.construx.com/Page.aspx?nid=14 http://www.se-radio.net/
Software \ Data Mining	http://www.bredemeyer.com/ http://www.construx.com/Page.aspx?nid=14 http://www.se-radio.net/ http://www-01.ibm.com/software/rational/ http://www.sei.cmu.edu/research/ http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm http://www.thearling.com/ http://databases.about.com/od/datamining/Data_Mining_and_Data_Warehousing.htm http://www.web-datamining.net/ http://www.tdan.com/view-articles/5263/
Software \ Algorithm	http://www.softpanorama.org/Algorithms/index.shtml http://cgm.cs.mcgill.ca/~godfried/teaching/algorithms-web.html http://www.cs.sunysb.edu/~skiena/214/lectures/ http://www.personal.kent.edu/~rmuhamma/Algorithms/algorithm.html http://java.sun.com/docs/books/tutorial/collections/algorithms/ http://www.analysttool.com/

	http://www.davidconsultinggroup.com/ http://www.microtool.de/e_index.htm http://www.offshoresoftwaresuccess.com/ http://www.reqbib.com/adavis/
HCI	http://www.pcd-innovations.com http://tc.eserver.org http://www.nathan.com http://www.hcibib.org http://nooface.net http://hci.stanford.edu/ http://www.si.umich.edu/msi/hci.htm http://www.sigchi.org/ http://www.cs.umd.edu/hcil/ http://hcibib.org/hci-sites/ http://dir.yahoo.com/Science/computer_science/human_computer_interaction__hci_/

Appendix C Instrument for Recommendation Evaluation

- Questions for each recommendation:

1. (This question ensures participants understand the page) Please identify, if any, some interesting sentences from the given page. You don't have to read the entire page. Please skim through headings, sections or sentences that seem interesting to you. You may click any links in the page if 1) they seem interesting to you or 2) they better help you to understand the information provided in the given page.

   After reading and understanding the page, do you agree with the following statements?
2. I feel that this recommendation is relevant to at least one topic of my interests – providing information related to the subject area(s) of my interests

   (1) Strongly Disagree … Strongly Agree (5)
3. I feel that this recommendation is novel -- providing new information to me and the information is beyond what I already know

   (1) Strongly Disagree … Strongly Agree (5)
4. I feel that this page is relevant to *the top 5 categorical topics of the page*

   (1) Strongly Disagree … Strongly Agree (5)

- Questions for interview:

1. Please tell me the reasons for your ratings of the above questions 2 and 3.
2. Please tell me which areas of computer science that you are interested in?

3. From a scale of 1 (not interested at all) to 5 (highly interested), how much are you interested in *a list of the system-predicted categorical interests for the user on the provided usage data*.

Appendix D：Diversity Index and Reported Interests

Table 0-1 displays each participant's DI value, identified interest clusters, and the self-reported user interests, ordered by DI values. Full names of cluster acronyms are listed above the table. The numbers following the cluster names indicate the size of the cluster. Larger numbers imply a stronger degree of user interest. The italic fonts in the "reported interests" columns are those system-predicted interests that were rated as 4 (agree) or 5 (strongly agree) by participants, and were not expressed in the self-reported interests.

AL: algorithm, AI: artificial intelligence, CA: computer architecture, CP: computer programming, CSawd: computer science award, CSlit: computer science literature, CSO: computer science organizations, CUR: currency, DB: databases, DS: data structure, GR: computer graphics, HCI: human computer interaction, MthOpt: mathematical optimization, OS: operating system, SE: software engineering, SR: security, ThCS: theoretical computer science.

Table 0-1 Diversity Index (DI), Clusters, and Self-Reported User Interests

Id	DI	Interest Clusters	Reported Interests
15	35	AL 4, DB 2, SE 2	compiler and programming language, AI, machine learning, data mining, IR *ThCS 4, AI 5, DB 4, AL 5*
14	38	AL 7, DB 2, CA 2, ThCS 2	cloud computing, virtualization, LLVM, OS *DS 5, CP 5, AL 4, HCI 4, AI 4*
1	50	SR 7, CSlit 2	information security-related things
24	56	OS 6, SR 4, CA 3, HCI 3, DB 2	N/A *HCI 5, SR 4, CA 5*
6	59	OS 7, DB 3, SR 2, CA 2	N/A *ThCS 5, OS 5, CP 5, DB 5, DS 4, SR 4, CA 4*
13	59	SR 5, CA 4, HCI 3, DB 2, OS 2	system and network administration, traffic

			monitoring, virtualiztion, programming, VOIP *AL 5, CSawd 5*
23	62*	AL 6, DB 3, SR 3, CA 2	computer networking* *SR 5, HCI 4, OS 5, DS 4, DB 4*
4	67	DB 3, OS 3, SR 2, CA 2, HCI 2	new technology, gadget, business related technology *DB 4, DS 5, CP4*
19	67	AL 3, GR 3, ThCS 3, DB 2, CP 2, DS 2	volume rendering, C++ language, distributed system, scientific applications *DB 4, CP 5, ThCS 5, GR 5, CSO 4*
16	68	DB 5, SR 3, AL 2, HCI 2, CP 2	db, asp, ajax, c# *DS 5*
25	68	AL 8, DB 3, ThCS 2, MthOpt 2	finance, mathematics, CS ThCS 4, CP 4, AL 4
10	70	SR 5, HCI 3, DB 2, CA 2, AI 2	Ecommerce, outsourcing, information technology, IT consulting *SR 4*
17	70	HCI 3, CA 3, GR 3, DB 2, OS 2	programming networking, beginning programming, mobile device programming, networking foundations *ThCS 5, HCI 5, DS 4*
3*	71	DB 5, SR 3, AL 2, CA 2	DBA, SQL, reporting service, BI, ETL *CP 3*
8	71	DB 5, AL 5, SR 3, CP 2	data mining, database structure *NA*
20	71	SR 4, DB 2, AL 2, HCI 2	databases, programming, security, web pages *HCI 5, GR 4*
5	76	AL 4, CA 4, DB 2, SR 2	software/program visualization, computer graphics, cell phone technology (not PDA)** *ThCS 5, AI 4,CP 5*
7	76	*SR 5, CSO	computer graphics, virtual reality, web programming *HCI 4, ThCS 4, SR 4, AI 4*
12	76	AL 4, CA 4, DB 2, SR 2	systems, networking, distributed algorithms, high performance computing, scheduling, computational complexity, computability theory *ThCS 5, CP 4, CUR 4*
2	78	AL 5, DB 2, SR 2, HCI 2, CP 2	data mining, databases, system design *ThCS 4, HCI 5, CP 4, DS 4, GR 4, AI 4, ML 4*
9	81	AL 4, HCI 3, SR 2, GR 2	GIS, Geovisulization, multivariate

			GR 4, HCI 4
18	81	DB 5, SR 3, AL 3, CA 2, CP 2	web application programming
22	81	*AL 4, DB 2, SR 2	flash game development, design, UI, server structure, QA, software engineering, and CS *GR 5, HCI 5, AI 4*
11	83	CA 4, AL 3, DB 2, SR 2, HCI 2, SE 2	news, daily tech, legalese, innovation, security, networking, IT *ThCS 4, HCI 4, GR 4, CP 4, DS 4, CSawd 5*
21*	91	SR 4, CP 3, HCI 2, GR 2, AI 2	computer games, surfing the internet *GR 4*

* discussed in section 4.3.2. ** for Id 5, our system recommended several PDA related pages.

Appendix E Informed Consent Form

You have been solicited as a participant in an experiment to evaluate a software system. This experiment is being conducted by Pei-Chia Chang, a graduate student in Information and Computer Sciences at University of Hawaii at Manoa, who will be glad to answer any questions you may have about the experiment.

There are two steps of the experiments - 1) providing certain web browsing history and 2) answering an online survey. In step 1, you will need to select and provide certain web browsing history log from Internet Explorer or Mozilla for the past month. You will only need to provide web visits to computer science related websites according to your judgment. Instructions are available at http://www2.hawaii.edu/~pcchang/ICS699/LogExtract.html Your provided data will be associated with your contact information in order to continue the 2nd step of the experiment. It takes about 5 - 25 minutes for you to extract the history data, depending on how much data you have. In step 2, you will need to response a survey about certain web pages, which will be selected based on the data you provided in step 1. Once you complete the 2nd step, your web browsing history and survey answers will be coded anonymously and confidentially for result analysis. The survey session is expected to last about 40 - 80 minutes, depending on the provided data in step 1.

There is no risk associated with this experiment. Although there is no official compensation for the participation, your help is greatly appreciated. Pei-Chia Chang will hold a thank you party in Hawaii with gifts at the end of the experiment.

As a participant, you have certain rights, which are listed below:

You have the right to withdraw from the experiment at any time for any reason.

If you decide to withdraw, please inform Pei-Chia Chang immediately. Otherwise, identification of your data might not be possible because of our efforts to ensure anonymity. If you decide to withdraw after completing step 1 and before starting step 2 of the experiment, your data will be guaranteed deleted. However, after the provided data will be coded in step 2, any identification will be lost. Lastly, at the completion of the experiment, all collected data will be destroyed. You may review the analyzed summary of this study, if you so desire. However, you will not be able to identify your own data due to anonymity.

Finally, we appreciate your time and effort in this experiment. There is no right or wrong answer. The purpose is to evaluate design features for a software system.

1. Do you voluntarily agree to participate in the experiment after reading this consent form in its entirety?

Your selection below indicates that you have read this consent form in its entirety and that you voluntarily agree to participate.

____ Yes

____ No

Any question regarding answering this?

2. Please provide your contact information if you are willing to participate in the experiment.

(For corresponding only, no identify information required)

Thank you very much!!

Researcher:

Pei-Chia Chang, PhD. Candidate

Information & Computer Sciences

University of Hawaii at Manoa

http://www2.hawaii.edu/~pcchang

peichia.chang@gmail.com 808-2209701

This research project (CHS# 17068) is under the supervision of

Committee on Human Studies

University of Hawaii at Manoa

1960 East-West Road

Biomedical Building, Room B-104

Honolulu, HI 96822

Phone: 808.956.5007

Email: uhirb@hawaii.edu

Bibliography

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules in Large Databases.* In Proc. of the 20th International Conference on Very Large Data Bases, pp. 487-499.

Andreas, L. S., & Pericles, A. M. (2005). *Agent Intelligence Through Data Mining (Multiagent Systems, Artificial Societies, and Simulated Organizations)*: Springer-Verlag New York, Inc.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*: Addison Wesley.

Balabanovi, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Commun. ACM, 40*(3), 66-72.

Bates, M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review, 13*(5), 407-424.

Belkin, N. J. (1977). *A Concept of Information for Information Science.* University of London, United Kingdom -- London.

Belkin, N. J. (1978). Progress in Documentation: Information Concepts for Information Science. *Journal of Documentation, 43*(1), 55-85.

Belkin, N. J. (1997). User Modeling in Information Retrieval. On *Tutorial presented at the Sixth International Conference on User Modeling*

Belkin, N. J., Brooks, H. M., & Oddy, R. N. (1979). Representing and Classifying Anomalous State of Knowledge. In *Informatics 5: The Analysis of Meaning* (pp. 227-238): London, Aslib.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., et al. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics, 7*(3), 154-165.

Bowman, C. M., Danzig, P. B., Manber, U., & Schwartz, M. F. (1994). Scalable Internet resource discovery: research problems and approaches. *Commun. ACM, 37*(8), 98-107.

Brooks, C. H., & Montanez, N. (2006). *Improved annotation of the blogosphere via autotagging and hierarchical clustering.* In Proc. of the 15th international conference on World Wide Web, pp. 625-632.

Budzik, J., Bradshaw, S., Fu, X., & Hammond, K. J. (2002). *Clustering for opportunistic communication.* In Proc. of the 11th international conference on World Wide Web, pp. 726-735.

Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.). (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*: IOS Press.

Burke, R. (1999). *The Wasabi Personal Shopper: a case-based recommender system.* In Proc. of the sixteenth national conference on Artificial Intelligence and the eleventh Innovative Applications of Artificial Intelligence conference pp. 844-849.

Burke, R. (2000). Knowledge-Based Recommender Systems. In A. Kent (Ed.), *Encyclopedia of Library and Information Systems* (Vol. 69.32). New York: Marcel Dekker.

Burke, R. D., Hammond, K. J., & Young, B. C. (1996). *Knowledge-based Navigation of Complex Information Spaces.* In Proc. of the 13th National Conference on Artificial Intelligence, pp. 462-468.

Burke, R. D., Hammond, K. J., & Young, B. C. (1997). The FindMe Approach to Assisted Browsing. *IEEE Expert: Intelligent Systems and Their Applications, 12*(4), 32-40.

Cardie, C. (1997). Empirical Methods in Information Extraction. *Artificial Intelligence Magazine 18*(4), 65-79.

Castells, P., Fernandez, M., & Vallet, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Trans. on Knowl. and Data Eng., 19*(2), 261-272.

Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1997). *Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases.* In Proc. of the 23rd International Conference on Very Large Data Bases, pp. 446-455.

Chang, P. C., & Quiroga, L. M. (2009). *Using Wikipedia Content to Derive an Ontology for Modeling and Recommending Web Pages across Systems.* In Proc. of the ACM RecSys'09 Workshop on Recommender Systems & the Social Web, pp. 79-83.

Chang, P. C., & Quiroga, L. M. (2010a). *An Ontological User Modeling Approach Using Wikipedia's Content for Cross-System Personalization.* In Proc. of the IUI2010 Workshop on User Data Interoperability in the Social Web, pp. 30-33.

Chang, P. C., & Quiroga, L. M. (2010b). *Using Wikipedia's Categories to Study the Coverage of a User's Interests for a Web Page Recommender.* In Proc. of the Poster and Demo Adjunct Proceedings of the 18th International Conference on User Modeling, Adaption, and Personalization (UMAP 2010), pp. pp. 7 - 9.

Charniak, E. (1993). *Statistical Language Learning*: MIT Press.

Chen, L., & Sycara, K. (1998). *WebMate: a personal agent for browsing and searching.* In Proc. of the second international conference on Autonomous Agents, pp. 132-139

Chin, D. N. (1993). Acquiring user models. *Artificial Intelligence Review, 7*(3-4), 185-197.

Chin, D. N. (2007). Information Filtering, Expertise and Cognitive Load. In D. D. Schmorrow & L. M. Reeves (Eds.), *Foundations of Augmented Cognition* (pp. 75-83). Berlin, Germany: Springer

Chin, D. N., & Crosby, M. E. (2002). Introduction to the Special Issue on Empirical Evaluation of User Models and User Modeling Systems. *User Modeling and User-Adapted Interaction, 12*(2-3), 105-109.

Collins, M. J. (1999). *Head-Driven Statistical Models for Natural Language Parsing.* Ph. D., University of Pennsylvania, Philadelphia, United States.

Cost, S., & Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning, 10*(1), 57-78.

Dai, H., & Mobasher, B. (2002). *Using Ontologies to Discover Domain-Level Web Usage Profiles*. Paper presented at the 2nd workshop on Semantic Web Mining, PKDD.

Davies, J., Fensel, D., & Harmelen, F. v. (Eds.). (2003). *Towards the Semantic Web: Ontology-driven Knowledge Management*: John Wiley & Sons.

Debnath, S., Mitra, P., & Giles, L. (2005). *Automatic extraction of informative blocks from webpages.* In Proc. of the 2005 ACM symposium on Applied computing, pp. 1722-1726.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science, 41*(6), 391-407.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia Risks. *Commun. ACM, 48*(12), 152-152.

Dewey, M. (1876). Dewey Decimal Classification (DDC) System. Retrieved Feb. 21, 2009, from http://www.oclc.org/dewey/

Dolog, P., & Nejd, W. (2003). *Challenges and Benefits of the Semantic Web for User Modeling.* In Proc. of the Adaptive Hypermedia 2003 Workshop, pp. 849-852.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2nd Edition)*: Wiley-Interscience.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. (2003). *Stuff I've seen: a system for personal information retrieval and re-use.* In Proc. of the 6th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03), pp. 72-79.

Eirinaki, M., & Vazirgiannis, M. (2003). Web Mining for Web Personalization. *ACM Transactions on Internet Technology, 3*(1), 1-27.

Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). *SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process*. Paper presented at the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.

Engineers, I. o. E. a. E. (1990). In A. C. o. I. S. C. Glossaries (Ed.), *IEEE Standard Computer Dictionary*. New York: IEEE

Erdelez, S. (1997). *Information Encountering: A Conceptual Framework for Accidental Information Discovery.* In Proc. of the International Conference on Research in Information Needs, Seeking and Use in Different Contexts pp. p.412-421.

Fensel, D. (2003). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*: Springer-Verlag New York, Inc.

Frommholz, I., Mehta, B., Niederée, C., Risse, T., & Thiel, U. (2004). *Supporting Information Access in Next Generation Digital Library Architectures.* In Proc. of the First International Workshop of the EU Network of Excellence DELOS on Digital Library Architectures, pp. 49-60.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The Vocabulary Problem in Human-System Communication. *Commun. ACM, 30*(11), 964-971.

Garcia, E. (2006). Cosine Similarity and Term Weight Tutorial. Retrieved Feb. 19, 2009, from http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html

Gauch, S., & Futrelle, R. (1993). *Experiments in Automatic Word Class and Word Sense Identification for Information Retrieval.* In Proc. of the Symposium on Document Analysis and Information Retrieval, pp. 425-434.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental Models*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature, 438*(7070), 900-901.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM, 35*(12), 61-70.

Gonzalez, G., Angulo, C., Lopez, B., & Rosa, J. L. d. l. (2005). *Smart User Models, Modeling the Humans in Ambient Recommender Systems.* In Proc. of the 10th International Conference on User Modeling, pp. 11-20.

Greaves, M., Holmback, H., & Bradshaw, J. (2000). What Is a Conversation Policy? In *Issues in Agent Communication* (pp. 118-131): Springer-Verlag.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud., 43*(5-6), 907-928.

Guarino, N., Masolo, C., & Vetere, G. (1999). OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems, 14*(3), 70-80.

Hammond, K. J. (1989). *Case-based planning: viewing planning as a memory task.*

Boston, Massachusetts, United States: Academic Press Professional, Inc.
Heckmann, D., & Krueger, A. (2003). A User Modeling Markup Language (UserML) for Ubiquitous Computing. In *User Modeling 2003* (pp. 148).
Heckmann, D., Schwartz, T., Brandherm, B., & Kröner, A. (2005). *Decentralized User Modeling with UserML and GUMO.* In Proc. of the International Conference on User Modelling, pp. 61-65.
Heer, J., Card, S. K., & Landay, J. A. (2005). *prefuse: a toolkit for interactive information visualization.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.
Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). *An algorithmic framework for performing collaborative filtering.* In Proc. of the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230-237.
Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst., 22*(1), 5-53.
Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). *Recommending and evaluating choices in a virtual community of use.* In Proc. of the Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 194-201.
Huberman, B. A., Pirolli, P. L. T., Pitknow, J. E., & Lukose, R. M. (1998). Strong Regularities in World Wide Web Surfing. *Science, 280*(5360), 95-97.
IEEE. (2000). Suggested Upper Merged Ontology. from http://www.ontologyportal.org/
Joachims, T., Freitag, D., & Mitchell, T. (1997). *WebWatcher: A Tour Guide for the World Wide Web.* In Proc. of the Fifteenth International Joint Conference on Artificial Intelligence, pp. 770-775.
Johnson-Laird, P. N. (1983). *Mental Models*: Cambridge University Press.
Joseph, S., Yukawa, J., Suthers, D., & Harada, V. (2006). *Searching Emergent Vocabularies: Exploring Methods to Reduce Cognitive Load during Web Navigation and Resource Contribution*. Paper presented at the 39th Annual Hawaii International Conference on System Sciences - Volume 07.
Kautz, H., Selman, B., & Shah, M. (1997). Referral Web: combining social networks and collaborative filtering. *Commun. ACM, 40*(3), 63-65.
Kim, H., Howland, P., & Park, H. (2005). Dimension Reduction in Text Classification with Support Vector Machines. *The Journal of Machine Learning Research, 6*, 37-53.
Kobsa, A. (1995). Editorial. *Using Modeling and User-Adapted Interaction, 4*(2

Special Issue on User Modeling Shell Systems), iii-v.

Kobsa, A. (2001). Generic User Modeling Systems. *User Modeling and User-Adapted Interaction, 11*(1-2), 49-63.

Kobsa, A. (2007). Privacy-enhanced Personalization. *Commun. ACM, 50*(8), 24-33.

Kolodner, J. (1993). *Case-based reasoning*: Morgan Kaufmann Publishers Inc.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM, 40*(3), 77-87.

Krulwich, B. (1997). Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. *Artificial Intelligence, 18,* 37-45.

Lin, D. (1998). *Using Collocation Statistics in Information Extraction*. Paper presented at the Seventh Message Understanding Conference (MUC-7).

Liu, G. Z. (1997). Semantic Vector Space Model: Implementation and Evaluation. *Journal of the American Society for Information Science, 48*(5), 395-417.

Maedche, A., Motik, B., Stojanovic, L., Studer, R., & Volz, R. (2003). Ontologies for Enterprise Knowledge Management. *IEEE Intelligent Systems, 18*(2), 26-33.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1995). *Discovering Frequent Episodes in Sequences*. Paper presented at the First International Conference on Knowledge Discovery and Data Mining.

Manning, C. D., & Schütze, H. (1999). Part-of-Speech Tagging. In *Foundations of Statistical Natural Language Processing* (pp. 340-380). Cambridge, MA: MIT Press. .

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Retrieved Feb. 21, 2009, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Maudlin, M., & Leavitt, J. (1994). *Web Agent Related Research at the Center for Machine Translation.* In Proc. of the ACM Special Interest Group on Networked Information Discovery and Retrieval.

Mehta, B., Niederée, C., & Stewart, A. (2005). *Towards Cross-system Personalization.* In Proc. of the 3rd International Conference on Universal Access in Human-Computer Interaction pp. 119-123.

Michael, J. P., & Daniel, B. (2007). Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization* (pp. 325-341): Springer-Verlag.

Middleton, S. E., Shadbolt, N. R., & Roure, D. C. D. (2004). Ontological User Profiling in Recommender Systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 54-88.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic Personalization Based on Web Usage Mining. *Commun. ACM, 43*(8), 142-151.

Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000). *Integrating Web Usage and Content Mining for More Effective Personalization.* In Proc. of the Proceedings of the First International Conference on Electronic Commerce and Web Technologies, pp. 165 - 176.

Mooney, R. J., & Roy, L. (2000). *Content-based book recommending using learning for text categorization.* In Proc. of the fifth ACM conference on Digital libraries, pp. 195-204.

Niederée, C. J., Stewart, A., Mehta, B., & Hemmje, M. (2004). *A Multi-Dimensional, Unified User Model for Cross-System Personalization.* In Proc. of the Workshop on Environments for Personalized Information Access, in Advanced Visual Interfaces International Working Conference, pp. 34-54.

Norman, D. A. (1983). Some Observations on Mental Models. In I. D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 241-244). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

O'Donnell, M., Mellish, C., Oberlander, J., & Knott, A. (2001). ILEX: an architecture for a dynamic hypertext generation system. *Nat. Lang. Eng., 7*(3), 225-250.

Oard, D. W. (1997). The State of the Art in Text Filtering. *User Modeling and User-Adapted Interaction, 7*(3), 141-178.

Oberle, D., Berendt, B., Hotho, A., & Gonzalez, J. (2003). *Conceptual User Tracking.* In Proc. of the Advances in Web Intelligence, pp. 155-164.

Ohkura, T., Kiyota, Y., & Nakagawa, H. (2006). *Browsing System for Weblog Articles based on Automated Folksonomy.* In Proc. of the WWW 2006.

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). *On the Inequality of Contributions to Wikipedia.* In Proc. of the 41st Annual Hawaii International Conference on System Sciences, pp. 304.

Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2001). *Personalization from Incomplete Data: What You Don't Know Can Hurt*. Paper presented at the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

Payne, S. J. (1991). A Descriptive Study of Mental Models. *Behavior and Information Technology, 10*, 3-21.

Pazzani, M., Muramatsu, J., & Billsus, D. (1996). *Syskill & Webert: Identifying Interesting Web Sites.* In Proc. of the Thirteenth National Conference on Artificial Intelligence, pp. 54-61.

Petrelli, D., Angeli, A. D., & Convertino, G. (1999). *A user-centered approach to user modeling.* In Proc. of the seventh international conference on User modeling,

pp. 255 - 264.
Pirolli, P. L., & Card, S. K. (1995, May 07 - 11, 1995). *Information Foraging in Information Access Environments.* In Proc. of the SIGCHI conference on Human factors in computing systems, pp. 51 - 58.
Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., et al. (2002). Personalized Search. *Commun. ACM, 45*(9), 50-55.
Plaisant, C., Grosjean, J., & Bederson, B. B. (2002). *SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation.* In Proc. of the Information Visualization, 2002. IEEE Symposium, pp. 57-64.
Ponzetto, S. P., & Strube, M. (2007). *Deriving a Large Scale Taxonomy from Wikipedia*. Paper presented at the 22nd National Conference on Artificial Intelligence.
Porter, M. F. (2001). Snowball: German Stemming Algorithm. Retrieved Feb. 19, 2009, from http://snowball.tartarus.org/texts/introduction.html
Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292-306.
Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., & Riedl, J. (2007). *Creating, destroying, and restoring value in wikipedia*. Paper presented at the 2007 international ACM conference on Supporting Group Work.
Quillian, R. (1968). Semantic Memory. In M. L. Minsky (Ed.), *Semantic Information Processing* (pp. 227-270). Cambridge, MA, United States: The MIT Press.
Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106.
Quintarelli, E. (2005). Folksonomies: Power to the People. Retrieved Feb. 21, 2009, from http://www.iskoi.org/doc/folksonomies.htm
Quiroga, L. M. (1999). *Personalized information organization: Acquisition and modeling of users' interest profiles in information filtering systems.* Ph.D. Dissertation, Indiana University, Indiana, United States.
Quiroga, L. M. (2009). *Using personal ontologies as a context based user modeling tool for bridging Information Filtering (IF) and Personal Information Management (PIM) research*. Paper presented at the Workshop on Context-Aware Recommender Systems
Quiroga, L. M., & Mostafa, J. (2002). An experiment in building profiles in information filtering: the role of context of user relevance feedback. *Inf. Process. Manage., 38*(5), 671-694.
Razmerita, L., Angehrn A., Nabeth, T. (2003). *On the role of user models and user modeling in Knowledge Management Systems*. Paper presented at the HCI

International.
Reichling, T., Schubert, K., & Wulf, V. (2005). *Matching Human Actors based on their Texts: Design and Evaluation of an Instance of the ExpertFinding Framework.* In Proc. of the 2005 international ACM SIGGROUP conference on Supporting group work, pp. 61-70.
Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). *GroupLens: an open architecture for collaborative filtering of netnews*. Paper presented at the 1994 ACM conference on Computer supported cooperative work.
Resnick, P., & Varian, H. R. (1997). Recommender systems. *Commun. ACM, 40*(3), 56-58.
Rich, E. A. (1979a). *Building and exploiting user models.* Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, PA, USA.
Rich, E. A. (1979b). User modeling via stereotypes. *Cognitive Science, 3*(4), 329-354.
Rich, E. A. (1999). Users are individuals: individualizing user models. *Int. J. Hum.-Comput. Stud., 51*(2), 323-338.
Richardson, L. (2005). The Ultra Gleeper: A Recommendation Engine for Web Pages. from http://www.crummy.com/software/UltraGleeper/IntroPaper.html
Riesbeck, C. K., & Schank, R. C. (1989). *Inside Case-Based Reasoning*: L. Erlbaum Associates Inc.
Rosenfeld, R. (1994). *Adaptive Statistical Language Modeling: A Maximum Entropy Approach.* Ph.D. Dissertation, Carnegie Mellon University.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523.
Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM, 18*(11), 613-620.
Saracevic, T. (1970). The Notion of 'Relevance' in Information Science. In T. Saracevic (Ed.), *Introduction to Information Science* (pp. 111-151): Bowker Co.
Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*(6), 321-343.
Schafer, J. B., Dan, F., Jon, H., & Shilad, S. (2007). Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization* (pp. 291-324): Springer-Verlag.
Schamber, L. (1994). Relevance and Information Behavior. *Annual Review of Information Science and Technology, 29*, 3-48.
Shani, G., & Gunawardana, A. (2009). *Evaluating Recommender Systems* (No. MSR-TR-2009-159): Microsoft Technical Report.

Shardanand, U., & Maes, P. (1995). *Social information filtering: algorithms for automating "word of mouth".* In Proc. of the Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 210-217.

Smith, G. (2004). Atomiq: Folksonomy: social classification. from http://atomiq.org/archives/2004/08/folksonomy_social_classification.html

Smyth, B., & McClave, P. (2001). *Similarity vs. Diversity.* In Proc. of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, pp. 347-361.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21.

Speer, R. (2007). *Open Mind Commons: An Inquisitive Approach to Learning Common Sense*. Paper presented at the Workshop on Common Sense and Intelligent User Interfaces, the International conference on Intelligent User Interfaces.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations Newsletter 1*(3), 12-23.

Stephen, P. H. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*. New Jersey: Princeton University Press.

Strube, M., & Ponzetto, S. P. (2006). *WikiRelate! Computing Semantic Relatedness Using Wikipedia.* In Proc. of the Artificial Intelligence, pp. 1419-1424.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). *YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia.* In Proc. of the International World Wide Web Conference, pp. 697-706.

Sussan, G. (2007). Web 2.0. In *The Academic Library and the Net Gen Student* (pp. 35): ALA editions.

Sutcliffe, R. F. E. (1991). *Distributed representations in a text based information retrieval system: A new way of using vector space model.* In Proc. of the 14th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 123-132.

Szomszor, M., Alani, H., Cantador, I., O'Hara, K., & Shadbolt, N. (2008). *Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis.* In Proc. of the 7th International Conference on The Semantic Web, pp. 632-648.

Toms, E. G. (2000). Understanding and facilitating the browsing of electronic text. *Int. J. Hum.-Comput. Stud., 52*(3), 423-452.

W3C. (2004). OWL Web Ontology Language. from

http://www.w3.org/TR/owl-features/
Wahlster, W., & Kobsa, A. (1989). User Models in Dialog Systems. In K. Alfred & W. Wolfgang (Eds.), *User models in dialog systems* (pp. 4-34): Springer-Verlag New York, Inc.
Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2004). *Text mining: predictive methods for analyzing unstructured information*. New York: Springer.
Wendlandt, E., & Driscoll, J. R. (1991). *Incorporating a semantic analysis into a document retrieval strategy.* In Proc. of the 14th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 270-279.
Williamson, K. (2002). Discovered by chance: The role of incidental information acquisition in an ecological model of information use *Library & Information Science Research, 20*(1), 23-40.
Wooldridge, M. (1999). Intelligent Agents. In G. Weiss (Ed.), *Multi-agent Systems: A Modern Approach to Distributed Artificial Intelligence* (pp. 27-78): The MIT Press.
Xu, Y., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol., 57*(7), 961-973.
Zhang, F., Song, Z., & Zhang, H. (2006). *Web Service Based Architecture and Ontology Based User Model for Cross-System Personalization.* In Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 849-852
Ziegler, C.-N., Lausen, G., & Schmidt-Thieme, L. (2004). *Taxonomy-driven computation of product recommendations.* In Proc. of the ACM international conference on Information and knowledge management, pp. 406-415.