國立交通大學

資訊管理研究所

碩士論文

一個基於新聞流與相關回饋之個人資料檔強化混合式的新聞推薦機制

A Profile-enhanced Hybrid News Recommendation Mechanism

based on News Flow and Relevance Feedback

研究生：黃元辰

指導教授：羅濟群 教授

中華民國九十九年六月

一個基於新聞流與相關回饋之個人資料檔強化混合式的新聞推薦機制

A Profile-enhanced Hybrid News Recommendation Mechanism based on News Flow and Relevance Feedback

研究生：黃元辰 Student: Yuan-Chen Huang

指導教授：羅濟群 Advisor: Chi-Chun Lo

國立交通大學

資訊管理研究所

碩士論文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Information Management

June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

一個基於新聞流與相關回饋之個人資料檔強化混合式的新聞推薦機制

研究生：黃元辰　　　　　　　　　　　　　　　　　　指導教授：羅濟群

國立交通大學資訊管理研究所

摘要

隨著電腦科技與網際網路的發展，資訊傳遞與交換的質與量都隨之俱增，為避免資訊過載造成決策時的錯誤判斷，推薦系統的發明與導入成為重要關鍵。在以內容為基礎的推薦系統，其推薦結果優劣除了受到所使用的演算法外，也受到個人資料檔品質的影響。此外個人資料檔雖能代表個人長期的喜好，卻無法反映出短期的興趣所在。

為此本論文提出一混合式推薦演算法，結合新聞流與相關回饋兩技術，藉此反應使用者短期與長期的喜好，以獲得較佳的推薦結果。本文並實作一新聞推薦網站，藉以驗證與比較不同演算法。經實作一新聞推薦網站進行實驗證實，本文所提出之混合式推薦演算法能較單一使用語意式內容基礎推薦方法將推薦精準度從 0.311 提升至 0.744，達到更佳的推薦結果供使用者使用。

關鍵字：推薦系統、混合式方法、新聞流、相關回饋、語意式內容基礎推薦

A Profile-enhanced Hybrid News Recommendation Mechanism based on News Flow and Relevance Feedback

Student: Yuan-Chen Huang Advisor: Chi-Chun Lo

Abstract

With the development of computer science and Internet, the exchange and delivery of information increases rapidly. To avoid decision mistake caused by information overload, recommendation system has been invented and introduced. For content-based recommendation system, the quality of recommendation result is affected not only by algorithm itself, but also the quality of user profile. Besides, a content-based recommendation system is unable to reflect short-term interest of users.

In this thesis we proposed a hybrid recommendation algorithm combined news flow and relevance feedback. With these two techniques, the system can reflect short-term and long-term user interests. We also implemented the algorithm in a news recommendation website, which helped us to validate our algorithm and compare to other algorithm. Through the experiment based on a news recommendation website, our algorithm has been proven that the hybrid algorithm performs better than semantic content-based recommendation algorithm which enhanced precision from 0.311 to 0.744, and provides better recommendation result to users.

Keywords: Recommendation System, Hybrid Algorithm, News Flow, Relevance Feedback, Semantic Content-based Recommendation

誌謝

能夠順利完成這份論文並且能夠完成碩士班的學業，首要感謝我的父母與兄嫂和侄子，他們給了我一個溫暖的家庭，讓我在外衝刺學業時，不論遇到什麼挫折與困難，隨時都能夠有一個避風港，讓我拋開一切煩惱。

接著我要感謝我的指導教授羅濟群老師，交大六年的師生情緣，老師帶給我的不是隻字片語能夠形容，感謝老師給予我在課業上的諄諄教誨，以及生活上的各種照顧，讓我在新竹的日子裡能夠從裡至外的蛻變。

再來要感謝實驗室裡的鼎元學長、志華學長與銘家學長，他們在課業上、計畫上的各種指導、幫助與討論，讓我能夠有機會不斷的汲取各種新知並且茁壯成長，同時了解做研究應有的態度與精神，使我進行碩士論文時能夠游刃有餘的面對各種問題。

然後我還要感謝實驗室的學長姐—栩嘉學姊、俊傑學長、邦曄學長、盈蓉學姊、芝蓉學姊、昌民學長、家偉學長，以及同學們—世豪、致衡、冠儒、志健、湘婷，他們在學業上與生活上帶給我的點點滴滴。當然也少不了學弟妹們—哲豪、秉賢、冠廷、光禹、慕均、棉媛、孟儒、芳儀、靜蓉，他們帶給我各式各樣的歡樂氣氛，讓我在苦悶的研究生涯中依舊能保持愉快的心情。

最後，我要感謝我的女朋友—雅婷。沒有妳這本論文可能無法順利完成，沒有妳我的生活將失去色彩，沒有妳我的情感便會枯竭，沒有妳我的人生便不能稱做完滿，千言萬語的感謝也無法表達我心中的激動，真的十分感謝妳。

Table of Contents

1896

List of Figures

List of Tables

List of Equations

ES A 1896

Chapter 1 Introduction

1.1 Research Background and Motivation

With the developing of computer, it really changes our live. Computer involves our daily life in many places, including science, education, finance, traffic... etc. We use computer to control, monitor, edit, and lots of works, which generates many information. In 1990s, the raising of WWW made people enable to exchange information faster than ever, and computer was been used almost in everywhere. With this circumstance, people create and exchange tons and pounds information in everyday. Take media as an example, in traditional we order a newspaper every day, receive a magazine every month, watch TV three hours in a day; now we open computer and browse the Internet, we can receive hundreds or even thousands of news which comes from ten or more media. This is quite different to old time, and computer does really change human's life.

However, with so much information how people digest information became a serious problem. If a person receives too many information in a short period, "information overload" may occur and lower her or his judgment. This issue was proposed for a very long time, while WWW was not created. And with the development of the Internet, it happens more and more often than before which people starts to do research for solving the problem [1-2].

To solve information overload, computer scientists induct “recommendation system” to discover explicit and implicit needs of human being and filtering unnecessary information [3-4].

Recommendation system has two types: content-based and collaborative filtering. The content-based recommendation comes from information retrieval community

[3-5]. Text documents are recommended based on a comparison between their content and a user profile, which using features extracted from the text of documents.

The collaborative filtering approach is quite different to content-based approach [3, 6-11]. Instead of recommend items similar to a user has liked in the past, collaborative filtering approach recommend items that similar users have liked. Based on this philosophy, collaborative filtering computes the similarities between users instead of items.

However, content-based recommendation system encounters several drawbacks: 1. with immature user profile, the system may have lower performance. 2. Profile is lack of change, which the recommendation result may not reflect truly desire of user after a while. A user may have highly interesting in some important event news even though he has no interesting originally, for example, the news of bankrupt of a country to a user have no interesting in finance. Also, at the beginning of setting profile, a user may not sense her or his real interesting. The potential interesting only can be observed by her or his reading.

1.2 Approach

In this thesis, we proposed a hybrid algorithm to enhance recommendation result fitting short-term and long-term interesting. The designed experiment platform, Top Story website, gave an example of the algorithm and also collected experiment data to validate it. After experiment, we proved that the hybrid algorithm has better performance than semantic content-based recommendation, which had been proven as a good recommendation method [12].

1.3 Thesis Outline

The remainder of this thesis is built as follow. In chapter 2, we gave a lecture review of related knowledge and technologies, such recommendation system and semantic web. In chapter 3, we proposed a concept called “News Flow”, which is based on “Knowledge Flow”. Also, we proposed a hybrid recommendation mechanism based on news flow and semantic content-based recommendation. In chapter 4, we described details of experiment and its result, and it also included a discussion about the result. In chapter 5, we gave a conclusion and suggestions for future work about this thesis.

Chapter 2 Related Works

In this thesis, we proposed a hybrid algorithm to solve low performance of recommendation result caused by insufficient profile. The necessary research background and relevant technologies includes: (1) Semantic web (SW), (2) Recommendation system, (3) Knowledge flow, (4) Relevance feedback. We will introduce them in the following sections

2.1 Semantic Web

In 1994 at the 1st International World Wide Web Conference, Tim Berners-Lee first mentioned the concept of semantic web [13-14]. In this article, Berners-Lee mentioned the need for semantics in the Web. The web is a set of nodes and links. To a user, this has become an exciting world, but there is very little machine-readable information there. With this situation, it had need to adding semantic meaning for the web. Adding semantics to the web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading.

Later Berners-Lee published “Semantic Web Road Map” on the Internet [15], and it was the first time semantic web were proposed officially. The core of semantic web is that through adding metadata for documents on the Internet, these documents would not only be understood by human beings but also be reasoned and processed by computers. He proposed Resource Description Framework (RDF) as the metadata [16].RDF is a standard model for data interchange on the Web. It has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

In 2001 Berners-Lee published “The Semantic Web” in the magazine – Scientific American [17], which expounded the meaning and future of semantic web. It described the evolution of a Web consisted largely data and information for computers to manipulate. In this article, he introduced “ontology” into the semantic web. With ontology, computer was more capable to handle the lexical and semantic meaning in the web and completed the semantic web. It is used to reason about the properties of that domain, and may be used to describe the domain.

Ontology originates from the philosophy of traditional Greek which is a branch of “Metaphysics”. It mainly focuses on categories of being and their relations. However, most ontology in modern society was implemented in computer science nowadays. In 1993, Gruber gave a strict definition for ontology [18]. He thought “*An ontology is an explicit specification of a conceptualization*.” Ontology in computer science is to find classes and objects in a given list, which represents concept and entity of objects, and describes their property, restriction, disjoint statement, and relation. Gruber also mentioned five rules of designing ontology:

1. Clarity: An ontology should effectively communicate the intended meaning of defined terms. Definitions should be objective.
2. Coherence: An ontology should be coherent: that is, it should sanction inferences that are consistent with the definitions. At the least, the defining axioms should be logically consistent.
3. Extendibility: An ontology should be designed to anticipate the uses of the shared vocabulary. It should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically.
4. Minimal encoding bias: The conceptualization should be specified at the

knowledge level without depending on a particular symbol-level encoding.

5. Minimal ontological commitment: An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities.

In 2006, Shadbolt et al with Berners-Lee published “The Semantic Web Revisited” [19], reviewed the development of semantic web and introduced tolls, techniques, and insights about the semantic web . In this article two issues were mentioned which are data integration and virtual uptake. Data integration is being achieved in large part through the adoption of common conceptualizations referred to as ontologies. In the past years, ontologies has been implemented in biology, medicine, genomics, and related fields. The origin ways about data exposure are HTTP, HTML, and XML. But uptake requires increasing the amount of data exposure in RDF.

In this thesis, the main recommendation algorithm is based on semantic content-based recommendation, which we focused on semantic expansion and will introduce in the following section.

2.1.1 Semantic Expansion

Semantic expansion combines two parts: query expansion (QE), and semantic web. In 1983, both Smeaton et al [20]and Yu et al[21] used statistical relations to expand query vectors, which the relations are easily generated from the document at hands. However, Peat et al found there are limitations to the effectiveness one can expect from such system [22]. In 1997, Pollitt introduced a system – HIBROWSE with query expansion by combining terms from different facets interactively to refine the query [23]. In 2003, Yee et al introduced Flamenco hierarchical browsing interface which allows users adding or removing facets while browsing a web image database and dynamically generating previews of query results [24]. Another way of query expansion is thesaurus-based QE, which is employing different thesaurus relationships

[25].

2.1.2 Ontology and Spreading Activation Model

In 2008, Gao et al proposed an approach based on ontology and spreading activation model [26]. The recommender system compares the collected data to similar data collected from others and calculates a list of recommended items for the user. Through combining the user ontology and spreading activation model, the capability of discovering of user's potential interests is enhanced.

Spreading activation model is proposed in 1975 by Collins et al in order to simulate human comprehension through semantic memory [27]. It reviewed the original spreading-activation theory developed by M. R. Quillian while trying to correct some common misunderstandings concerning it [28]. It extended the theory in several respects, showed how the extended theory dealt with recent experimental findings, and compared it to the model of Smith, Shoben, and Rips [29].

Spreading activation model is an organization structure of long-term memory in human brain. Crestani et al used spreading activation model in information retrieval to expand the search vocabulary and to complement the retrieved document sets [30]. It established a prototype Web search system that exploits the differences between documents usually managed by IR systems and the Web.

In 2005, Aswath et al presented an automated, high precision-based information retrieval solution to boost item findability by bridging the semantic gap between item information and popular keyword search phrases [31]. A two level spreading activation network activates and hence identifies strong positive and negative phrases related to the matches of a given keyword search phrase, which in turn activates other potentially relevant products in addition to those that are exact keyword matches for the search term itself. Next, a SVM classifier is trained, using these strong positive and negative matches of a search phrase, to separate the rest of the matches from mismatches.

In 2008, Weng et al combined ontology and spreading activation model to develop a research paper recommendation system [32]. It proposed to use ontology and the spreading activation model for research paper recommendation that it can elevate the performance of the recommendation system and also improve the shortcomings of today's recommendation systems. This study utilized ontology to construct user profiles and makes use of user profile ontology as the basis to reason about the interests of users. Furthermore, it took advantage of the spreading activation model to search for other influential users in the community network environment, making a study on their interests in order to provide recommendation on related information.

Cantador et al also published a thesis in 2008 which combined above methodologies with context-aware for recommend news [12, 33]. They established a News@hand system combined content features and collaborative information to make news suggestions. Item and user profile are represented in terms of concepts appearing in domain ontologies. The semantic relations among these concepts are exploited to enrich the above representations and incorporated within the recommendation processes. Besides, they also introduce context-aware into its recommendation method. Context-aware makes the system able to sense the environment of user and expand the query, which enhances the recommendation result.

2.2 Recommendation System

Recommendation system has two approaches: content-based and collaborative filtering recommendation [3], which has quite different philosophy in their methodologies. We introduce them in the following sections.

2.2.1 Content-based Recommendation

The content-based recommendation comes from information retrieval community. Text documents are recommended based on a comparison between their content and a user profile, which using features extracted from the text of documents. Some weighting formulas were introduced and gave higher weights to discriminating words.

Pure content-based recommendation has many implementations, such as InfoFinder [34], News Weeder [35], and TREC [36]. Pure content-based approach has several disadvantages. First, only a shallow analysis of certain kinds of content can be supplied. If we faced some domain such as movies or music, there is no amendable extraction method to implement content-based approach. Even with text documents like web pages, sometimes information retrieval techniques would ignore subjective qualities, multimedia information, and network factors.

Second, if the system can only recommend items highly against a user's profile, the result would be restricted to user's profile. User would see recommendation results similar to those already read. A randomness of recommendation is able to deal the problem [3].

Finally, there is still a common problem to most recommendation systems that is eliciting user feedback. Rating documents is a burden to users. With fewer ratings requires better qualities. With the pure content-based approach, a user's own ratings are the only factor influencing future performance. If the quality of ratings goes bad, the result will be a mess.

2.2.2 Collaborative Filtering Recommendation

The collaborative filtering approach is quite different to content-based approach. Instead of recommend items similar to a user has liked in the past, collaborative filtering approach recommend items that similar users have liked [3, 7-8]. Based on this philosophy, collaborative filtering computes the similarities between users instead of

1896

items. Scores for unseen items are predicted based on a combination of score known from the nearest neighbors. A pure collaborative filtering recommendation does no analysis of the items but sees as an identifier. Recommendations will come from users have similar identifiers.

In 1994, Konstan et al proposed the internet news recommendation system – GroupLens [11], which helps people find articles they will like in the huge stream of available articles. The system collected ratings from Usenet readers and used those ratings to predict how much other readers would like an article before they read it. This recommendation engine was one of the first automated collaborative filtering systems in which algorithms were used to automatically form predictions based on historical patterns of ratings.

In 1995, Hill et al introduced the Bellcore video recommender [37], which helps people in choosing multimedia in the domain of music and musical artists. It contains three interfaces: filtering, recommendation, and prediction interface, which aims to evaluate the power of a particular form of virtual community to help users find things they will like with minimal search effort.

Besides user based collaborative filtering algorithm, item-to-item collaborative filtering algorithm was proposed by Vucetic et al in 2000 [38]. It proposed an alternative regression-based approach that searches for relationships among items instead of looking for similarities among users. Later, Amazon.com implemented this algorithm in their e-commerce business [9]. Rather than matching the user to similar users, item-to-item collaborative filtering matches each of the user’s purchased and rated items to similar items, then combines those similar items into a recommendation list. With the list, once a user purchased any product in the list, the system recommends other products in the similar group to user.

Pure collaborative filtering recommendation solves all disadvantages given for pure content-based approach. By using other's recommendations, we can deal with any kind of content and receive items with dissimilar content to those seen in the past. And it maintains effective performance even with fewer ratings from any individual user [3, 7, 10].

But still, it has its own disadvantages. If a new item comes into the database, there is no way to be recommended until more information about it is obtained through another user either rating it or specifying which other items it is similar to. Therefore, if the number of users is small relative to the volume of information in the system, than there is a danger of the coverage of rating becoming very sparse, thinning the collection of recommendable items. The second issue is if a user's taste is quite different to others, it will lead to poor recommendation performance. These two problems depend on the size and composition of the user population, which also influence a user's group of nearest neighbors.

2.3 Knowledge Flow

With the exchange of knowledge among people, it generates flows in an organization. The exchange of knowledge had become an important issue in knowledge management. To recognize knowledge constitutes a valuable intangible asset for creating and sustaining competitive advantages [39]. Knowledge sharing activities are generally supported by knowledge management systems. Thus, knowledge flow became popular which is able to help people understand how knowledge exchange and transfer among people in a group.

In different researches, people gave different definitions for knowledge flow. In 2002, Fung et al said knowledge flow is the use of patent citation data between firms in the same industry or different industries [40]. They measured knowledge flows by calculating overlaps between firms and internal intensity of knowledge flows within an industry.

In 2002, Zhuge said a knowledge flow is a process of knowledge passing between people or knowledge processing mechanism [41]. It has three crucial attributes:

1. Direction: determine the sender and the receiver.
2. Content the sharable knowledge content.
3. Carrier: the media that can pass the content.

Zhuge uses an arrow to denote the direction of the knowledge flow. The carrier can be based on the Internet or a local network. The sharable knowledge content means the knowledge is understandable by all members of team.

Anjewierden et al defined knowledge flow in 2005 which in weblogs is a communication pattern where the post of one blogger links to that of another blogger to exchange knowledge [42]. They also chose some relevant areas which knowledge flow can be of value:

1. Monitoring the “frequency” and “intensity” of crucial flows between people in and outside of the organization.
2. Evaluating the content of the crucial flows between organizational entities to detect bottleneck and emerging problems.
3. Monitoring the development of flows over time to keep track of developments in the knowledge household.

The concept of knowledge flow came from knowledge management, whose goal is enhancing the effectiveness of teamwork by accumulating and sharing knowledge among team members to facilitate peer-to-peer knowledge sharing [41]. To improve the efficiency, Zhuge proposed a pattern-based approach that combines codification and personalization strategies in order to design an effective knowledge flow network [43]. In 2003, Kim et al proposed a knowledge flow model combined with a process-oriented approach to capture, store, and transfer knowledge [44]. Anjewierden proposed that a KF forms a communication pattern whereby the post of one blogger links to that of another blogger to exchange knowledge in 2005 [42].In 2008, Luo et al proposed textual knowledge flow based on a semantic link network for the discovery of knowledge innovation, intelligent browsing and personalized recommendation in Web services and e-Science Knowledge Grid [45].

In 2009, Lai et al thought knowledge workers normally have various information needs over time when performing tasks [46]. Thus, they define a knowledge flow from the perspective of a worker's information needs to represent the evolution of referencing behavior and the knowledge accumulated for a specific task.

2.4 Relevance Feedback

In the typical information retrieval environment, including both ad-hoc interactive retrieval and document, filtering based on long-term information needs, an original query is submitted to a system which then returns documents for inspection. Users then look at those retrieved documents and submit a new query based upon their original need and the returned documents. Relevance feedback is the process of automatically altering an existing query using information supplied by users about the relevance of previously retrieved documents.

Relevance feedback has been an important research topic for well over 15years. Increased attention has been paid to relevance feedback in the past several years due to both the increased acceptance of statistical information retrieval systems which can easily use relevance feedback, and the effects of the TREC conferences. Evaluation of relevance feedback on the small pre-TREC test collections was notoriously difficult. The TREC routing environment offers a straightforward context in which relevance feedback can be evaluated and compared.

Thus, relevance feedback inputs the user's judgments on previously retrieved documents to construct a personalized query. These algorithms utilize the distribution of terms over relevant and irrelevant documents to re-estimate the query term weights, resulting in an improved user query [47-48].

The theory of relevance feedback is well-developed in vector space model [49] and also in the binary independence probabilistic model. In 1976, Robertson et al introduced relevance feedback into binary independence probabilistic model and were able to produce a partial ranking of the retrieved set of documents [50]. They extended the research and were able to produce a full ranking of the retrieved set of documents [51-52].

Due to expansion of recommendation system implementation, relevance feedback is also implemented in image retrieval. Deselaers et al proposed a new method for relevance feedback in image retrieval [53]. They also proposed a scheme to learn weighted distances which can be used in combination with different relevance feedback methods.

1896

Chapter 3 Hybrid Recommendation Mechanism based on News Flow and Relevance Feedback

In chapter 3, we identify our problem first in Section 3.1. And we will describe our design rationale in Section 3.2. Therefore, in Section 3.3 we will introduce the overview of proposed recommendation mechanism. Then for each step of mechanism, we will introduce in Sections 3.4 to 3.8. In the end of chapter, we have discussions in Section 3.9 to talk about possible advantages and disadvantages of proposed mechanism.

3.1 Problem Definition

In this thesis, our main goal is to improve recommendation result that can reflect short-term and long-term information need. Traditional content-based recommendation algorithm recommends items based on user profile. However, sometimes user may have interests on something she/he doesn't like. Also, long-term interest may change which occurs difference between user profile and interest. Thus, traditional recommendation algorithm may have inaccurate recommendation result, and that is our focused issue.

3.2 Design Rationale

The design rationale of proposed mechanism is surrounding to reflect short-term and long-term interesting of information need. Short-term interest may not reflect on user profile due to that is not user's original habits. But it can be observed by user's recent reading history. If a user read some news related to the topic she/he doesn't like, that may hint that she/he has interests in the topic recently. With this assumption, for short-term interest we design our mechanism that observed past reading events in a week to find what topic of news user may like. For long-term part, user’s interest should reflect on her/his reading history. Therefore we introduced relevance feedback to adjust user profile.

3.3 Overview of Recommendation Mechanism

In this thesis, we proposed a hybrid recommendation algorithm which combines both semantic content-based recommendation and news flow. Therefore, we will introduce the concept architecture of the algorithm in this section.

The algorithm has the following 5 steps:

1. Preprocessing news content
2. Calculating semantic content-based similarity score
3. Evaluating the news flow score
4. Calculating hybrid score
5. Adjusting user profile by relevance feedback

The detail information for each step will be introduced in Sections 3.4 to 3.8.

From the above steps, the major goal is to reflect short term interest that user profile may not represent. And for the long term habit changing, we implemented relevance feedback mechanism to adjust user profile.

3.4 Preprocessing

The first step of the algorithm is “preprocessing”, which is using information retrieval techniques to find representative terms of every document. Preprocessing includes the following 4 steps:

1. Tokenization
2. Stop word removing
3. Stemming
4. Evaluate terms

In this thesis, we use “Term Frequency – Inverted Document Frequency” (TF-IDF) to evaluate each word in the document. The TF-IDF can separate as two parts: (1) term frequency, and (2) inverted document frequency. We will introduce them in the later.

“Term frequency” is a simple measurement to evaluate a word in a given document. It is just count the term i showing times divided by total words in the document j. The equation is in the following:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

“Inverted document frequency” is a way to evaluate a word’s weight in given documents. The more documents the word shows, the less important the word is. It is the total number of documents divided by number of documents which the term is shown, and then we take the logarithm of that quotient. The equation is in the following:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad (2)$$

After we got TFs and IDFs, we multiply them and as the weight of each term. For term i in the document j, the TF-IDF equation is in the following:

$$\left(tf\text{-}idf\right)_{i,j} = tf_{i,j} \times idf_i \qquad (3)$$

3.5 Semantic Content-based Recommendation

Semantic content-based recommendation algorithm includes two steps:

1. Semantic expansion
2. Calculate cosine similarity

In this section, we will introduce how we implements semantic content-based recommendation as a foundation part of recommendation algorithm. Semantic Content-based Recommendation has been proved as a good recommendation algorithm, which we will implement it as control group. Basically, it is based on two parts: "Semantic expansion", and "Vector space model". We will introduce them in the following sections.

1896

3.5.1 Semantic Expansion

Semantic expansion is a way using ontology, which describes attributes of object itself and relations of between objects, to implement query expansion. With ontology, computer can understand which related items are able to be added into query set.

In this thesis, we implemented "Semantic closeness expansion algorithm". The algorithm uses an initially empty set of semantically close terms (SC terms). Link traversal has an associated cost, so closeness values degrade as expansion continues, until a cutoff threshold is reached. The algorithm is shown in the following:

```
initialTerm.Closeness = 1
cutoffThreshold = 0

initialTerm.Expanded = false
add initialTerm to SCTerms

while (any SCTerms have not been expanded)
    for (each unexpanded SCTerm)
        for (each link to another term)
            calculate traversalCost for this link type
            newValue = sourceTerm.Closeness - traversalCost
            if (newValue >= cutoffThreshold) then
                if (SCTerm contains targetTerm) then
                    if (targetTerm.Closeness < newValue ) then        (4)
                        targetTerm.Closeness = newValue
                        targetTerm.Expanded = false
                    else begin
                        targetTerm.Closeness = newValue
                        targetTerm.Expanded = false
                        add targetTerm to SCTerms
                    end if
                end if
            end if
        end for
        sourceTerm.Expanded = true
    end for
end while
```

Here we give an example to explain semantic expansion. For a term “Tigers”, in an ontology about sports we are able to discover that it is an instance of class “team”. There are some classes which has relations with the class, such as “player” or “league”. With the relations, we are able to find some instance in “player” like “Fu-Te Ni”, “Verlander” and in “league” like “MLB”. These terms can be added into query set, which users might have interests in news related to these terms.

3.5.2 Vector Space Model

Vector space model is an algebraic model for representing text documents as vectors of identifiers. Documents and queries are represented as vectors. For example:

$$\vec{d_j} = \left(w_{1,j}, w_{2,j}, \cdots, w_{t,j} \right)$$
$$\vec{q} = \left(w_{1,q}, w_{2,q}, \cdots, w_{t,q} \right)$$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. In this thesis what we used is TF-IDF weighting.

With vector space model, relevance of documents are able to be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as same kind of vector as the documents. To calculate two documents or queries which represented in vector, we can calculate their cosine of the angle as the similarity:

$$\cos\theta = \frac{\vec{d}, \vec{q}}{\left\|\vec{d}\right\| \left\|\vec{q}\right\|}$$

A cosine value of zero means that the query and document vector are orthogonal and have no match, which we see the two documents has no similarity between them.

3.6 News Flow

"News flow" is a major concept that we proposed in this thesis. The term is originally from mass communication, which means similar set pattern of news coverage [54]. In this thesis, it represents a sequence of historic news related to a single event. With news flow, in our proposed algorithm we used news flow to reflect reader's short term interest which user profile might not contain.

In this thesis, we collected news flow via doing cluster analysis. Clustering is a method of unsupervised learning. With clustering, we would get related news groups which help us to calculate the news flow score. In this thesis, we use hierarchical clustering algorithm with average-link to find news flow. The hierarchical clustering algorithm is the following steps:

```
SimpleHAC($d_1, \cdots, d_N$)
for n $\leftarrow$ 1 to $N$
    do for $i \leftarrow$ 1 to $N$
        do $C[n][i] \leftarrow Sim(d_n,\ d_i)$
        $I[n] \leftarrow 1$
$A \leftarrow []$
for $k \leftarrow$ 1 to $N$ -1
    do$(i,\ m) \leftarrow \text{argmax}_{\{\langle i,m\rangle : i\neq m \wedge I[i]=1 \wedge I[m]=1\}}$                (5)
        $A$.append$(\langle i,m\rangle)$
        for $j \leftarrow$ 1 to $N$
            do C$[i][j] \leftarrow Sim(i,\ m,\ j)$
               C$[j][i] \leftarrow Sim(i,\ m,\ j)$
        $I[m] \leftarrow 0$
return $A$
```

In the previous algorithm, the distance between documents is the inverse of the similarity between documents. The similarity is based on Vector Space Model (VSM), which the equation is in the following:

$$sim\left(d_i, d_j\right) = \frac{\overrightarrow{v_{di}} \cdot \overrightarrow{v_{dj}}}{\left\|\overrightarrow{v_{di}}\right\| \left\|\overrightarrow{v_{dj}}\right\|} \qquad (6)$$

With news flow from previous step, we are able to calculate the news flow score. For a given news flow, its score is based on news been read. To avoid older reading event affect the result. The score will multiply a time factor tw, which is in the following equation:

$$t\omega = \frac{7-d+1}{7} \quad (7)$$

In this thesis, the unit of time weight factor is measured in days. It is because news generates and reader reads in everyday. If the algorithm is applied to other domain such as academic paper recommendation, the unit can be replaced by month or season. 7 means days of week that is our focused period length, so if the algorithm is applied to other domain, 7 will be replaced to focused period length.

Besides, we use r to represents reading event for a user. It would be 1 if user read the news or 0 if not. Its equation is in the following:

$$r_j : \begin{cases} \text{if news is read, } r_j = 1 \\ \text{if news is not read, } r_j = 0 \end{cases} \quad (8)$$

With the previous two factors, the score for the news flow k is evaluated in the following equation:

$$nfs_k = \frac{\sum_{j=1}^{J} \left(r_j \bullet \frac{2}{\frac{1}{sim\left(up_i, d_j\right)} + \frac{1}{tw}} \right)}{\sum_{j=1}^{J} r_j}, \; d_j \in f_k \qquad (9)$$

Here we give another example to explain news flow score. Assume a person who has been read 15 news articles in the past week. Through clustering analysis, we can discover that they belong to three news flow. We calculate the similarities of 15 news articles, and then we can get the following result:

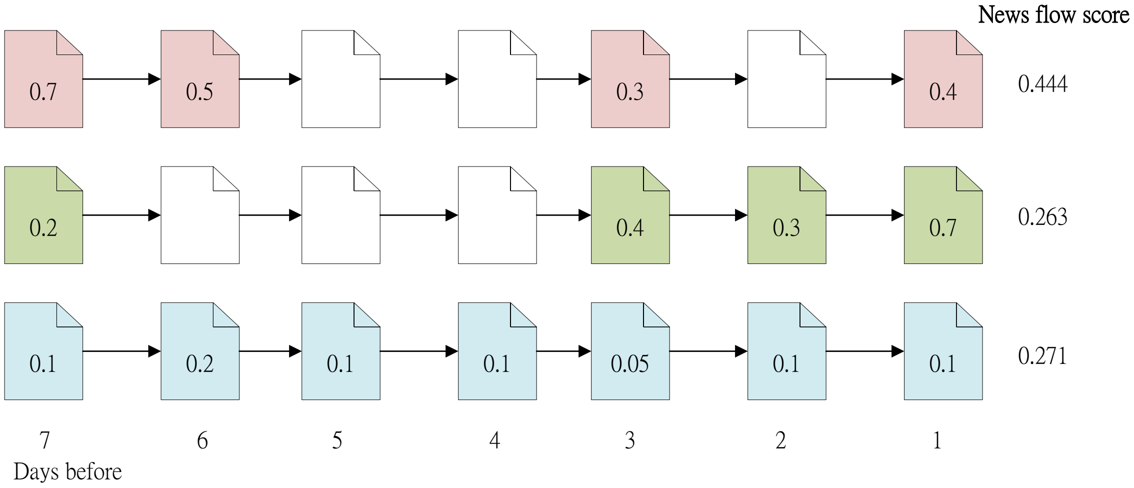


Figure 1. Example of News Flow Score

1896

For every number in the document symbol, it means the similarity of the news. From the previous figure, we can discover on average the user has lower interest in blue news flow with her or his profile. However, her or his reading behavior showed she or he has strong interest in the topic. With our algorithm, the news flow score recovers the blind spot which increased the weight of reading times about a flow. With more reading times, the blue news flow has a higher score than the green one. Based on the result, the system are possible to recommend some news about the blue news flow.

3.7 Hybrid Score

From the previous steps, we will have two scores for single news: one from semantic content-based recommendation algorithm; the other from the news flow which it belongs to. The score from semantic content-based recommendation represents a traditional way to value recommendation result based on user profile. And the score from news flow represents another way to value recommendation result based on past reading event. In this thesis, we propose a hybrid score based on these two score. The equation is in the following:

$$HS_{dj} = \alpha \cdot nfs_k + (1-\alpha) \cdot sim(up_i, d_j), \ 0 \le \alpha \le 1 \qquad (10)$$

In this equation, α represents the weight of news flow. The higher the value of α is, the more important the news flow is. In this thesis, α value will be tested in the experiment. The result will be shown in Chapter 4.

3.8 Relevance Feedback

In this section we will introduce relevance feedback mechanism as a way to reflect long term habit change. A user's interest might change while time passed. With a user profile that set in a month or season ago might lead bad recommendation result. If we based a good recommendation algorithm, it promised we would get good result matched the user profile. However, if a user's truly interest was no longer match to her or his profile, it will reflect on user's reading behavior. If a user does not like an issue or topic anymore, the related recommendation news will be ignored or read in a very short time. On the other hand, if a user does like an issue or topic, she or he will read the recommendation news in a reasoning time period. Based on previous steps, user would get a collection of news based on her or his profile. If the system supervised user's reading, it can review its recommendation result, which can also be implemented to adjust its profile and fit user's real interest. With this concept, we introduced relevance feedback mechanism into the system.

While user is reading news articles, the system will monitor her/his reading. With these information, the system would know which news user likes and do not like. Thus, to adjust her or his profile, we developed our formula based on Rocchio Classification formula [47, 49]. The adjustment is based on the following equation:

$$\begin{array}{l} p(t_i):\begin{cases}\text{if } w_i+\delta\le 1,\ w_i=w_i+\delta \\ \quad\text{if } wi+\delta>1,\ w_i=1\end{cases} \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad ,\ 0\le\delta\le 1 \\ n(t_i):\begin{cases}\text{if } w_i-\delta\ge 0,\ w_i=w_i-\delta \\ \quad\text{if } w_i-\delta<0,\ w_i=0\end{cases} \\ \quad t_i\subseteq T:\{t_1,t_2,\cdots,t_n\},\ tf\text{-}idf_i\ge tf\text{-}idf_{i+1}\end{array} \tag{11}$$

The variable δ represents the degree of adjustment, which would also be tested in experiment to show its effect. w_i represents the weight of term i in profile. In this thesis, we chose $t_i\in\{t_1,t_2,\cdots,t_{20}\}$ as the set of adjusted terms. If a news article was recognized as a positive reading, the system would parse the document and find important terms to execute relevance feedback. On the other hand, if it was a negative reading such as read time less than 5 seconds or much than 3 minutes, the term weight would decrease δ value.

3.9 Discussions

Based on our design rationale, our proposed mechanism is suitable for document recommendation due because reading items may have characteristic of continuity and sequence. In this thesis we designed the mechanism suited for news recommendation. In order to implement in other type of document recommendation, Equation 7 should change to suitable time measurement. Hence the unit of day should change to month or season due to the frequency of new document entered, and the value of 7 should change to a suitable short-term period.

On the other hand, if the recommendation item has no characteristic of continuity and sequence such as foods or drinks, our proposed mechanism is not suitable for the type of recommendation.

Chapter 4 Experiment Platform and Result Analysis

In this chapter, we will introduce our experiment platform – Top Story, a news recommendation website. Also, we will introduce the detail of experiment setup and result.

4.1 Experiment Platform – Top Story Website

To testify our algorithm, we established a news recommendation website called “Top Story”. The reasons we chose website as platform are: 1. easily to access, 2. easily collecting relevance feedback. The major system architecture is in the following figure:

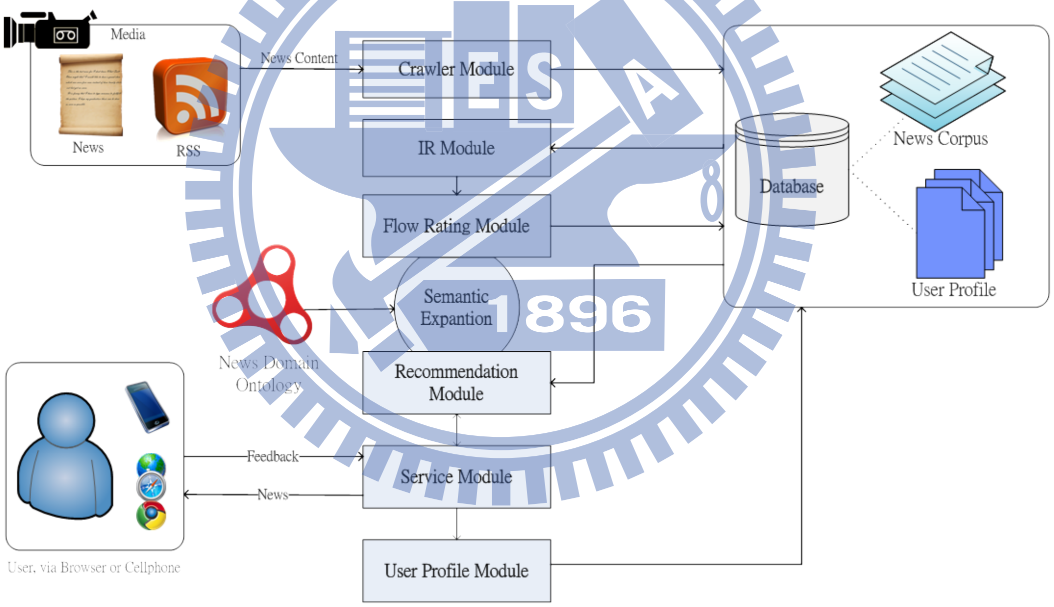


Figure 2. System Architecture

Crawler module collects news based on RSS every day. In our experiment, we collected news from three media sources: The Guardian, British Broadcasting Corporation, and Reuters.com. The crawler parsed news webpage and stored into database including news title, author, date, content, and publisher. The information will provide to user when they use the system.

IR module charges information retrieval techniques for preprocessing. This module would do tasks including tokenization, removing stop words, stemming, TF-IDF calculating. These tasks would be completed after getting the news.

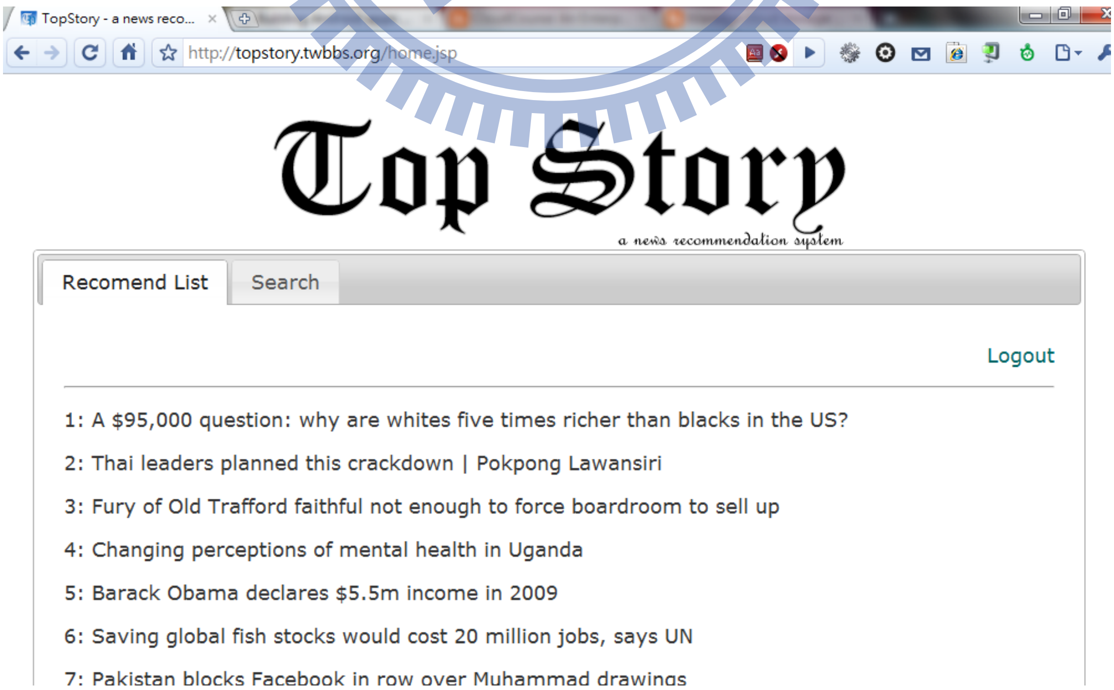
Flow rating module does clustering analysis for news item after doing the previous steps. The system would be able to get relative news in a cluster which we called “news flow”.

Recommendation Module calculates each daily news score plus news flow score; ranks them and generates a recommendation list.

Service Module handles web pages interacting with users. This module uses recommendation lists the previous module generates to provide news content to users. Also, it collects relevance feedbacks come from users and stores to the database for the resting use.

User profile module uses relevance feedbacks the previous module collects to adjust user profile with Equation 11 in chapter 3.

The system user interface is in the following figure:


TopStory - a news reco...
http://topstory.twbbs.org/home.jsp
Top Story
a news recommendation system
Recomend List
Search
Logout

1: A $95,000 question: why are whites five times richer than blacks in the US?

2: Thai leaders planned this crackdown | Pokpong Lawansiri

3: Fury of Old Trafford faithful not enough to force boardroom to sell up

4: Changing perceptions of mental health in Uganda

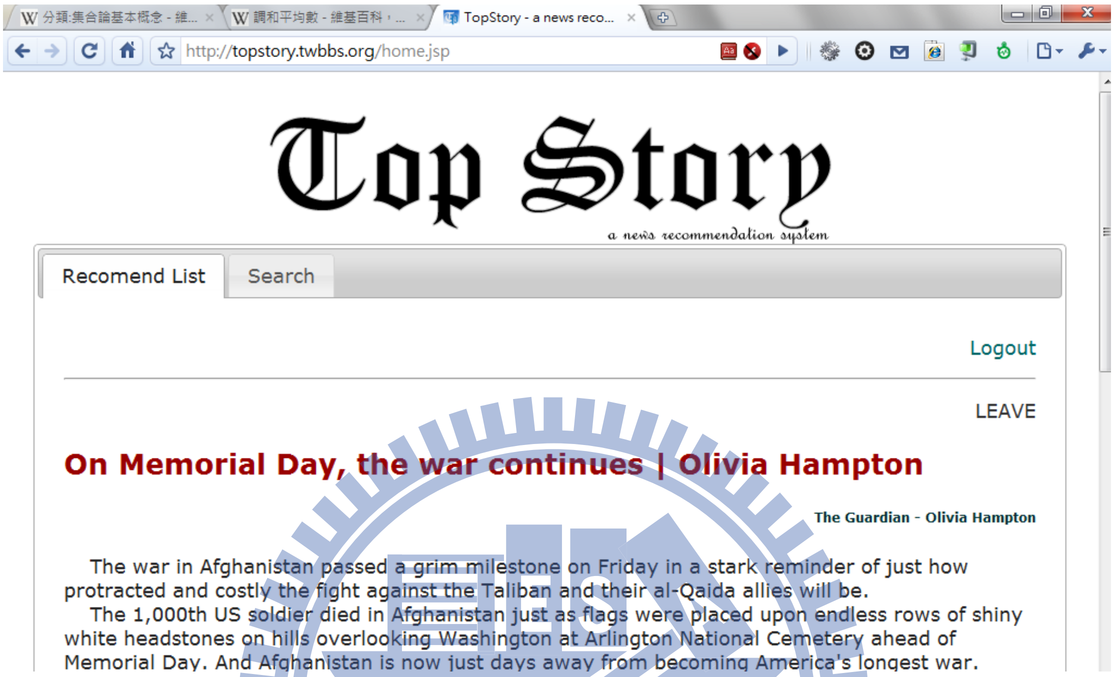
5: Barack Obama declares $5.5m income in 2009

6: Saving global fish stocks would cost 20 million jobs, says UN

7: Pakistan blocks Facebook in row over Muhammad drawings

Figure 3. System User Interface

When a user logins into the system, this webpage provides 20 news articles to the user. The user can choose any news she or he has interest, then continue the read in the following user interface:


分類:集合論基本概念 - 維...
調和平均數 - 維基百科，...
TopStory - a news reco...
http://topstory.twbbs.org/home.jsp
Top Story
a news recommendation system
Recomend List
Search
Logout
LEAVE

On Memorial Day, the war continues | Olivia Hampton

The Guardian - Olivia Hampton

The war in Afghanistan passed a grim milestone on Friday in a stark reminder of just how protracted and costly the fight against the Taliban and their al-Qaida allies will be.

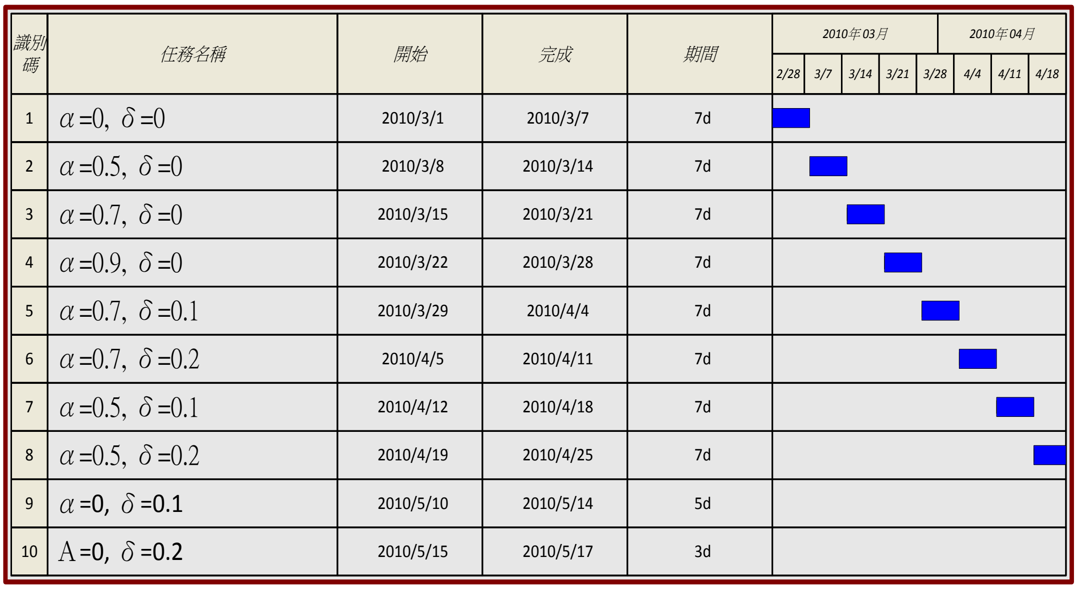
The 1,000th US soldier died in Afghanistan just as flags were placed upon endless rows of shiny white headstones on hills overlooking Washington at Arlington National Cemetery ahead of Memorial Day. And Afghanistan is now just days away from becoming America's longest war.

Figure 4. User Interface for Reading News

The Top Story system officially begins at March 1st, 2010, and keeps running until the experiment is end. The system was designed to be available for public, which the population of user contains any person from the Internet.

4.2 Experiment Setup

To collect enough experiment data and test two factors, α and δ, we separate the experiment into 10 periods. Each value of α and δ are shown in the following Gantt chart:



識別碼	任務名稱	開始	完成	期間	2010年03月					2010年04月		
					2/28	3/7	3/14	3/21	3/28	4/4	4/11	4/18
1	α=0, δ=0	2010/3/1	2010/3/7	7d	■							
2	α=0.5, δ=0	2010/3/8	2010/3/14	7d		■						
3	α=0.7, δ=0	2010/3/15	2010/3/21	7d			■					
4	α=0.9, δ=0	2010/3/22	2010/3/28	7d				■				
5	α=0.7, δ=0.1	2010/3/29	2010/4/4	7d					■			
6	α=0.7, δ=0.2	2010/4/5	2010/4/11	7d						■		
7	α=0.5, δ=0.1	2010/4/12	2010/4/18	7d							■	
8	α=0.5, δ=0.2	2010/4/19	2010/4/25	7d								■
9	α=0, δ=0.1	2010/5/10	2010/5/14	5d								
10	A=0, δ=0.2	2010/5/15	2010/5/17	3d								

Figure 5. Experiment Gantt Chart

The first 4 periods mainly tested α value. In these periods, δ keeps value of 0, and we tested 4 different value of α, which tried to find the best value for the mechanism.

The later 4 periods we picked 2 values of α which have better performance. With the two values of α, we testified the dominate value of δ. Each period continued for a week to ensure collecting enough data. And the last 2 periods is for testifying different values of δ under $\alpha = 0$.

4.3 Experiment Result

In the thesis, the major performance matrix is precision. The definition of precision is: “the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search”. Or we can formulate as the following equation 12:

$$Precision@N = \frac{\left|\{relevant\ documents\} \cap \{Top\ N\ retrieved\ documents\}\right|}{\left|\{Top\ N\ retrieved\ documents\}\right|} \qquad (12)$$

Precision is also evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at *n* or $P@n$. In this thesis, we considered n as 20 due to our comparable group [12] is using 5, 10, and 15.

The result data is shown in the following table:

Table 1. Experiment Result

Sampling Period	**Performance matrix**	**Precision@20**	**Avg. Reading Time(sec)**
Mar. 1st to Mar. 7th	α = 0; δ = 0	0.3111	48.6484
Mar. 8th to Mar. 14th	α = 0.5; δ = 0	0.6000	54.9292
Mar. 15th to Mar. 21st	α = 0.7; δ = 0	0.6308	72.2224
Mar. 22nd to Mar. 28th	α = 0.9; δ = 0	0.5333	80.5454
Mar. 29th to Apr. 4th	α = 0.7; δ = 0.1	0.7444	87.2126
Apr. 5th to Apr. 11th	α = 0.7; δ = 0.2	0.6000	83.0540
Apr. 12th to Apr. 18th	α = 0.5; δ = 0.1	0.6250	100.3973
Apr. 19th to Apr. 25th	α = 0.5; δ = 0.2	0.5250	125.9144
May 10th to May 14th	α = 0; δ = 0.1	0.3667	108.4098
May 15th to May 17th	α = 0; δ = 0.2	0.2929	75.2521

※Note:

Sampling Period: The period to collect experiment sampling data
P@20: Precision at 20, the precision of the top 20 recommendation items

To compare these data clearly, we use bar charts to express the data, which are shown in the following:

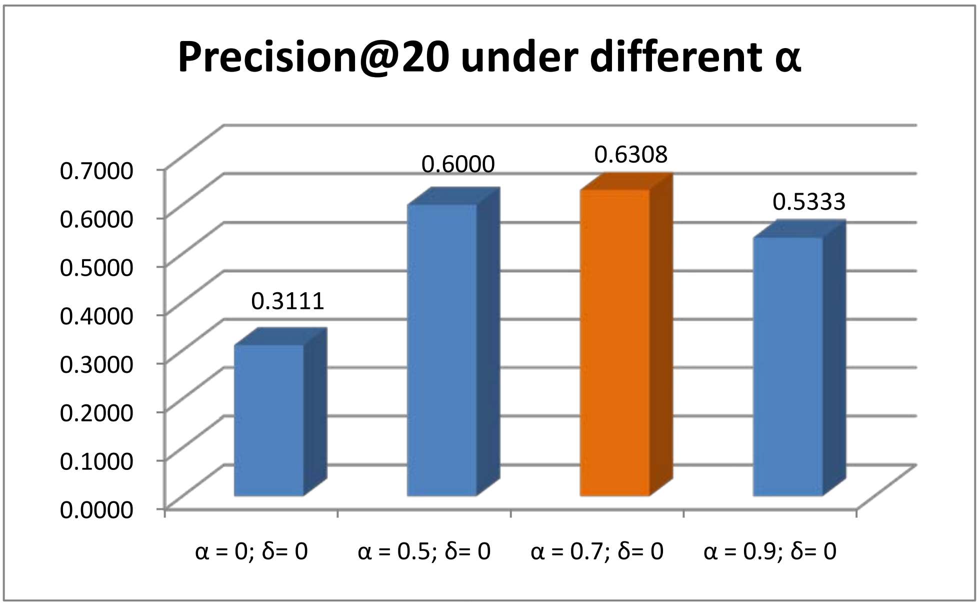


Figure 6. Precision under different α

In this bar chart, we compared precisions under different α value. The result shows if we implemented recommendation system under $\alpha = 0$, which means only implementing semantic content-based recommendation, we received a poor performance. Therefore, we tested hybrid algorithm with 3 different α value. The precisions with 20 recommendation items were 0.6, 0.63, 0.53 which under $\alpha = 0.5, 0.7, 0.9$.

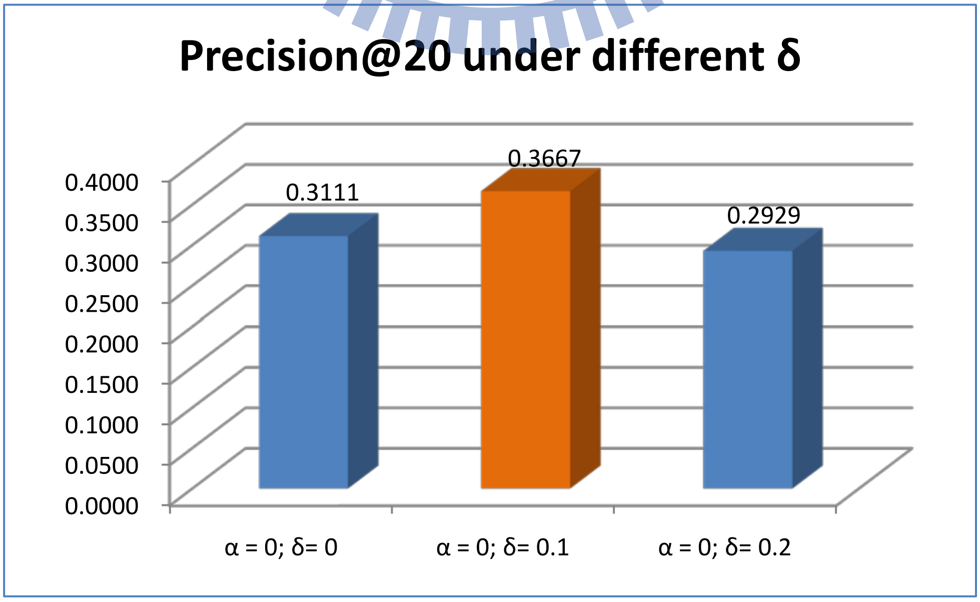


Figure 7. Precision under different δ

Figure 7 shows precisions under different δ values. The δ value determines the effect of relevance feedback. Based on the bar chart, we can figure out that if we set δ value to 0.1, the precision will increase. However, if we increase δ value to 0.2, the precision will go down and worse than original precision.

With the two experiment results, we mixed two variables to testify the performance of our proposed hybrid algorithm. We picked $\alpha = 0.5, 0.7$ and went through the experiment under $\delta = 0,\ 0.1, 0.2$. The result shows in the following chart:

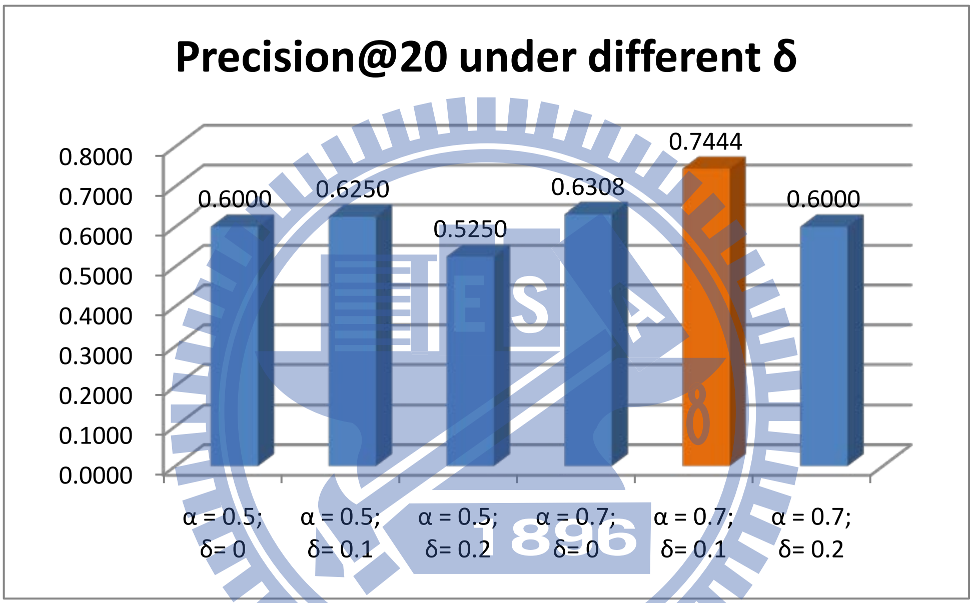


Figure 8. Precision under different δ and $\alpha = 0.5, 0.7$

In this bar chart, we observed that with different α value, $\delta = 0.1$ still received the best performance.

With the previous results, we compared two systems, one is our system, the other is News@Hand [12], a news recommendation which only implemented semantic content-based recommendation with context-aware. The comparison chart is in the following:

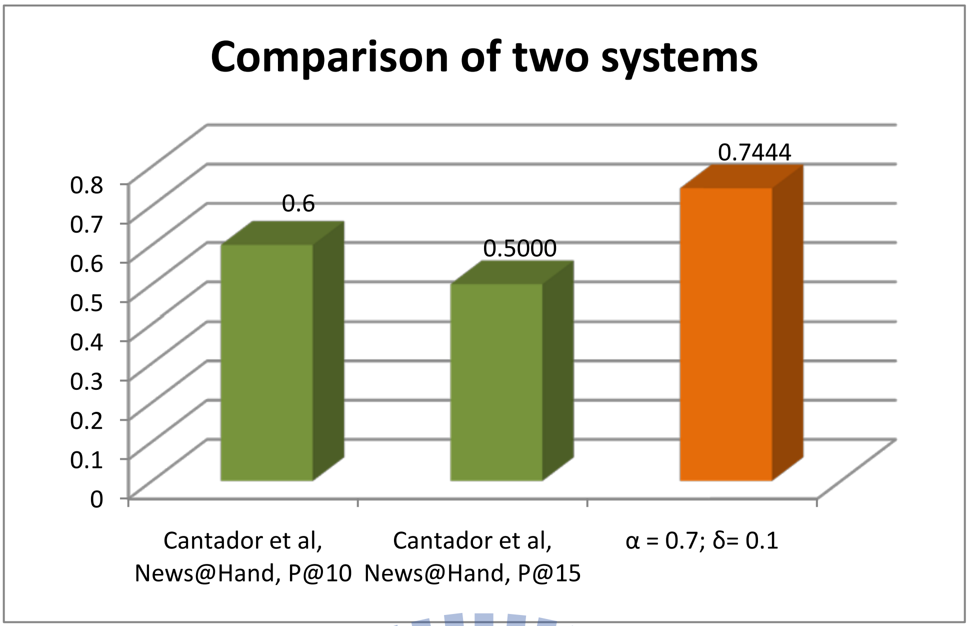


Figure 9. Comparison of two systems

From the previous bar chart, we could figure that even a good recommendation algorithm that had been proven would receive poor performance if recall goes up. With news flow, recommendation results were able to reflect implicit short-term interesting which most recommendation algorithms were ignored.

In conclusion, our experiment results show the proposed algorithm has better performance compared to non-hybrid algorithms. The best variable set is $\alpha = 0.7$ and $\delta = 0.1$. With news flow and relevance feedback, the precision increased from 0.311 to 0.744. The α value has the most significant effect to the whole recommendation performance which increased precision to 0.6308, and the δ value enhanced a part of performance and made performance better.

Chapter 5 Conclusion and Future Work

This thesis emphasizes issues about user profile in content-based recommendation. In this chapter, we summarize our studies and discuss our possible future work.

5.1 Summary

In this thesis, our main goal is proposing and validating an algorithm which can reflect implicit needs of news that user profile is unable to show. Therefore, we proposed a hybrid algorithm combined news flow and semantic content-based recommendation. In chapter 4, we proved our hybrid algorithm had a better performance than single-way algorithm. The proposed algorithm was able to improve recommendation performance.

5.2 Future Work

In the future, we can switch to collaborative filtering recommendation to replace original semantic content-based recommendation. Collaborative filtering recommendation has different recommendation philosophy to the content-based recommendation, which may have a better performance than content-based one.

Also, we can add context-aware into semantic content-based recommendation part. In our lecture review, we surveyed semantic content-based combined context-aware would have better performance. With context-aware, content-based recommendation can do better query expansion, which is proved to enhance recommendation performance. The two proposals can be implemented into the hybrid algorithm and testified their performances

References

[1] G. A. Miller, “Information Input Overload,” *In proceeding of Conference on Self-Organizing System*, 1962.

[2] S. R. Hiltz, and M. Turoff, “Structuring Computer-Mediated Communication Systems to avoid Information Overload,” *Communications of the ACM*, vol. 28, no. 7, pp. 680-689, 1985.

[3] M. Balabanovic, and Y. Shoham, “Fab: Content-Based, Collaborative Recommendation,” *Communications of the ACM*, vol. 40, no. 3, pp. 66-72, March 1997, 1997.

[4] A. Ansari, S. Essegaier, and R. Kohli, “Internet Recommendation Systems,” *Journal of Marketing Research*, vol. 17, pp. 363-375, 2000.

[5] Y. Wang, N. Stash, L. Aroyo *et al.*, “Semantic relations for content-based recommendations,” *In proceedings of the fifth international conference on Knowledge capture*, pp. 209-210, 2009.

[6] J. A. Konstan, B. N. Miller, D. Maltz *et al.*, “GroupLens: applying collaborative filtering to Usenet news,” *Communications of the ACM*, vol. 40, no. 3, pp. 77-87, 1997.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen *et al.*, “Evaluating collaborative filtering recommender systems,” *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5-53, 2004.

[8] J. B. Schafer, D. Frankowski, J. Herlocker *et al.*, “Collaborative Filtering Recommender Systems,” *Lecture Notes in Computer Science*, vol. 4321, pp. 291-324, 2007.

[9] G. Linden, B. Smith, and H. York, “Amazon.com recommendations: item-to-item collaborative filtering,” *IEEE Internet Computing*, vol. 7, no. 1, pp. 76-80, 2003.

[10] X. Su, and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Advances in Artificial Intelligence*, 2009.

[11] P. Resnick, N. Iacovou, M. Suchak *et al.*, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” *In proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.

[12] I. Cantador, A. Bellogin, and P. Castells, “Ontology-based Personalised and Context-aware Recommendations of News Items,” *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 562-565, 2008.

[13] T. Berners-Lee, “Plenary at WWW Geneva 94,” *http://www.w3.org/Talks/WWW94Tim/Overview.html*, 1994.

[14] T. Berners-Lee, R. Cailliau, A. Luotonen *et al.*, “The World-Wide Web,” *Communications of the ACM*, vol. 37, no. 8, pp. 76-82, 1994.

[15] T. Berners-Lee, “Semantic Web Road map,” *http://www.w3.org/DesignIssues/Semantic.html*, 1998.

[16] D. Brickley, “The WWW Proposal and RDF,” March 2001.

[17] T. Burners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American Magazine*, vol. 284, no. 5, pp. 34-43, 2001.

[18] T. Gruber, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing,” *International Journal of Human-Computer Studies*, vol. 43, no. 5, 6, pp. 907-928, 1993.

[19] N. Shadbolt, T. Berners-Lee, and W. Hall, “The Semantic Web Revisited,” *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96-101, 2006.

[20] A. F. Smeaton, and C. J. v. Rijsbergen, “The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System,” *Computer Journal*, vol. 26, pp. 239-246, 1983.

[21] C. T. Yu, C. Buckley, K. Lam *et al.*, “A Generalized Term Dependence Model in Information Retrieval,” *Information Technology: Research and Development*, vol. 2, pp. 129-154, 1983.

[22] H. J. Peat, and P. Willett, “The limitations of term co-occurence data for query in document retrieval systems,” *Journal of the American Society for Information Science*, vol. 42, no. 5, pp. 391-407, 1991.

[23] A. Pollitt, “Interactive information retrieval based on facetted classification using views,” *In Proceeding 6th International Study Conference on Classification*, pp. 51-56, 1997.

[24] K.-P. Yee, K. Swearingen, K. Li *et al.*, “Faceted metadata for image search and browsing,” *Conference on Human Factors in Computing Systems*, pp. 401-408, 2003.

[25] D. Tudhope, C. Binding, D. Blocks *et al.*, “Query expansion via conceptual distance in thesaurus indexed collections,” *Journal of Documentation*, vol. 62, no. 4, pp. 509-533, 2006.

[26] Q. Gao, J. Yan, and M. Liu, “A Semantic Approach to Recommendation System based on User Ontology and Spreading Activation Model,” *IFIP International Conference on Network and Parallel Computing*, pp. 448-492, 2008.

[27] A. M. Collins, and E. F. Loftus, “A Spreading-Activation Theory of Semantic Processing,” *Psychological Review*, vol. 82, no. 6, pp. 407-428, 1975.

[28] A. M. Collins, and M. Quillian, “Retrieval time from semantic memory,” *Journal of Verbal Learning and Verbal Behavior*, vol. 8, pp. 240-247, 1969.

[29] E. E. Smith, E. J. Shoben, and L. J. Rips, “Structure and process in semantic memory: A featural model for semantic decisions,” *Psychological Review*, vol. 1, pp. 214-241, 1974.

[30] F. Crestani, and P. L. Lee, “WebSCSA: Web Search by Constrained Spreading Activation,” *In proceeding of the IEEE Forum on Research and Technology Advances in Digital Libraries*, pp. 163, 1999.

[31] D. Aswath, J. D'cunha, S. T. Ahmed *et al.*, “Boosting Item Keyword Search with Spreading Activation,” *In proceedings of the 2005 IEEE/WIC/ACM International Conference*, pp. 19-22, 2005.

[32] S.-S. Weng, and H.-L. Chang, “Using ontology network analysis for research document recommendation,” *Expert Systems with Applications: An International Journal*, vol. 34, no. 3, pp. 1857-1869, 2008.

[33] D. Vallet, P. Castells, M. Fernandez *et al.*, “Personalized Content Retrieval in Context Using Ontological Knowledge,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 336-346, 2007.

[34] B. Krulwich, and C. Burkey, “Learning user information interests through the extraction of semantically significant phrases,” *In proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[35] K. Lang, “NewsWeeder: Learning to Filter Netnews,” *In Proceedings of the 12th International Conference on Machine Learning*, pp. 331-339, 1995.

[36] D. Hull, and S. Robertson, “The TREC-8 Filtering Track Final Report,” *Proceedings of the TREC-8 conference*, 2000.

[37] W. Hill, L. Stead, M. Rosenstein *et al.*, “Recommending and evaluating choices in a virtual community of use,” *In Conference on Human Factors in CImputing Systems*, pp. 194-201, 1995.

[38] S. Vucetic, and Z. Obradovic, “A Regression-Based Approach for Scaling-Up Personalized Recommender Systems in E-Commerce,” *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*, 2000.

[39] A. Cabrera, and E. F. Cabrera, “Knowledge-Sharing Dilemmas,” *Organization Studies*, vol. 23, no. 5, pp. 687-710, 2002.

[40] M. K. Fung, and W. W. Chow, “Measuring the intensity of knowledge flow with patent statistics,” *Economics Letters*, vol. 74, no. 3, pp. 353-358, 2002.

[41] H. Zhuge, “A knowledge flow model for peer-to-peer team Knowledge sharing and management,” *Expert Systems with Applications*, vol. 23, no. 1, pp. 23-30, 2002.

[42] A. Anjewierden, R. d. Hoog, R. Brussee *et al.*, “Detecting Knowledge Flows in Weblogs,” *13th International Conference on Conceptual Structures (ICCS 2005)*, pp. 1-12, 2005.

[43] H. Zhuge, “Knowledge flow management for distributed team software development,” *Knowledge-Based System*, vol. 15, no. 8, pp. 465-471, 2002.

[44] S. Kim, H. Hwang, and E. Suh, “A process-based approach to knowledge-flow analysis: a case study of a manufacturing firm,” *Knowledge and Process Management*, vol. 10, no. 4, pp. 260-276, 2003.

[45] X. Luo, Q. Hu, W. Xu *et al.*, “Discovery of Textual Knowledge Flow Based on the Management of Knowledge Maps,” *Concurrency and Computation: Practice and Experience*, vol. 20, no. 15, pp. 1791-1806, 2008.

[46] C.-H. Lai, and D.-R. Liu, “Integrating knowledge flow mining and collaborative filtering to support document recommendation,” *The Journal of Sysetm and Software*, vol. 82, no. 12, pp. 2023-2037, 2009.

[47] I. Ruthven, and M. Lalmas, “A survey on the use of relevance feedback for information access systems,” *The Knowledge Engineering Review*, vol. 18, no. 2, pp. 94-145, 2003.

[48] M. J. Huiskes, and M. S. Lew, “Performance Evaluation of Relevance Feedback Methods,” *In proceedings of the 2008 international conference on Content-based image and video retrieval*, pp. 239-248, 2008.

[49] J. J. Rocchio, “Relevance Feedback in Information Retrieval,” *In The SMART Retrieval System*, pp. 313-323, 1971.

[50] S. E. Robertson, and K. S. Jones, “Relevance weighting of search terms,” *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129-146, 1976.

[51] S. E. Robertson, and S. Walker, “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval,” *In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 345-354, 1994.

[52] S. E. Robertson, and S. Walker, “Okapi/Keenbow at TREC-8,” *IN proceeding of the 8th Text Retrieval COnference*, pp. 151-162, 2000.

[53] T. Deselaers, R. Paredes, E. Vidal *et al.*, “Learning weighted distances for relevance feedback in image retrieval,” *International Conference on Pattern Recognition 2008*, 2008.

[54] S. Nash, “International News Flow,” *Pacific Journalism Review*, vol. 2, no. 1, pp. 50-57, 1995.

1896