# Using Web 2.0 To Design An English Article Recommendation System

**Chen-Chung Chi, Chin-Hwa Kuo, and Wen-Jui Yu**
Department of Computer Science and Information Engineering of Tamkang University, Taiwan, R. O. C
ryanjih@mail.tku.edu.tw, chkuo@mail.tku.edu.tw, 799410161@s99.tku.edu.tw

**Abstract -** The spirit of Web 2.0 is through the collective power of communities to create, share and comment on the opinions of the user and others. The purpose of this study is to use the concept of Web 2.0 to design an English article recommendation system for senior high school students who are reading and learning. Five different databases of English vocabularies are utilized in this work. First, IWiLL, SHSETs and Web News Corpus are used for data pre-processing. Then they are combined with the GEPT level six to do article difficulty calculation. Lastly, phrases of three words in the article are compared with PyDict dictionary to translate into Chinese for learners to remember and learn. In this study, the system is combined with Facebook; learners only need to use a Facebook account to log on to the system with the use of Facebook Social Plugins components so that learners who like the article may share and recommend it to friends on Facebook.

Index Terms - Web 2.0, Facebook, Recommendation System.

## 1. Introduction

With the trend of globalization, English is the most common international language and it is very important for non-English-speaking countries, such as Taiwan. In addition, with the rapid development of the Internet and smart devices, articles are readily available. However, it is not easy to find suitable articles for learning. Therefore, the purpose of this study is to find articles for high school students.

This section will introduce the relevant background knowledge and literature. First Web 2.0 will be introduced, followed by Facebook, the recommendation system, the Input Hypothesis, corpus, IWiLL, and the NLTK tools. The implementation of the system will use the concept of Web 2.0 and the aforementioned tools.

### *A. Web 2.0*

Web 2.0[1] was coined by O'Reilly Media Inc. in 2004 at a conference; it refers to trends in the evolution of the Web application, including the collective strength of the community to create, share and comment on the opinions of the user and others.

### *B. Facebook*

Facebook [2] is today the most people used social networking site, established by Harvard University student Mark Zuckerberg in 2004 and opened in 2007 for developers to use a variety of APIs to publish applications to the Facebook platform. In 2010, Open Graph and Graph API introduced the use of Social Plugins for access by external web sites.

### *C. Recommendation System*

Resnick & Varian (1997) [3] formally proposed that the recommendation system is a one-by-one guide mechanism for a user in an environment with overwhelming information. Schafer et al (2001) [4] thought the recommendation system comes from recorded user preferences in order to guide the user's behaviour by using the professional knowledge of experts or the consumer behaviour itself. Today's widely used recommendation system contains both content filtering and collaborative filtering to create a hybrid approach.

*1) Content Filtering:* Is often used when recommending articles by analyzing user preferences to establish a list of personal interests. The keyword set in the web content is matched with the list of user interests using a similarity calculation and content with a high similarity value is recommended to the user.

*2) Collaborative Filtering:* Information of interest is recommended to groups with similar interests, preferences, or experiences. User ratings on articles are recorded to do analysis and achieve the purpose of collaborative filtering by recommending articles suitable for users.

*3) Hybrid Approach:* The merge of two or more recommended approaches of recommendation system. Combine the advantages of the recommendation system to compensate for the shortcomings of the recommendation system used in the handling of specific issues.

### *D. The Input Hypothesis Model*

Krashen proposed the Input Hypothesis Model [5][6] in 1985. This theory by Input Hypothesis, Acquisition/Learning Hypothesis, Monitor Hypothesis, Natural Order Hypothesis, Affective Filter Hypothesis of five mutually connected "what if". The Input Hypothesis, Krashen advocates the input language cannot be too difficult or too easy.

### *E. Corpora*

*1)* GEPT(General English Proficiency Test) [7]: GEPT vocabulary library is a reference to British Colins Cobuild English Dictionary (Bands 2 ~ 5), the English vocabulary for senior high school education revised by the College Entrance Examination Center (CEEC) in the second half of 2004 (a total of six levels), and the level four/six English vocabulary from China. GEPT vocabulary library divides the English vocabulary into six levels with a total of 6604 English words/phrases. In this study, use GEPT vocabulary library to calculate difficulty value of the English article.

*2)* SHSETs (senior high school English textbooks): Text

of senior high school English textbooks in Taiwan. In this study, the Sanmin Press and East Press compilation of the high school English text was used as one of the recommended articles database.

*3)* Web News: Articles published on the English news site. In this study, articles published by the Taiwan News English news website [8] were used as one of the recommended articles database.

*4)* PyDict[9]: This is an English/Chinese Dictionary, and it contains 177,746 words. In this study, this is the database that is used to translate English words into Chinese.

### *F. IWiLL*

IWiLL(Intelligent Web-based Interactive Language Learning) [10] is designed for high school students and includes the "IWiLL Community" and "IWiLL Campus", with learning and teaching platforms for both environments. The IWiLL research team was formed in 1999 and as of May 2012, there are 364 participating schools, 1,996 teachers, and 156,989 students registered. There are over 5000 English articles published by registered students. In this study, the proficiency in English vocabulary of senior high school students is inferred from the articles published by the registered students at IWiLL

### *G. NLTK*

NLTK (Natural Language Toolkit) [11] is a cross platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. The functionality is quite powerful.

## 2. System Architecture

The system is divided into four major parts. It consists of the operation of the system login, collection of online news, data pre-processing, and the article difficult evaluation and grading. Figure 1 shows the system architecture diagram.

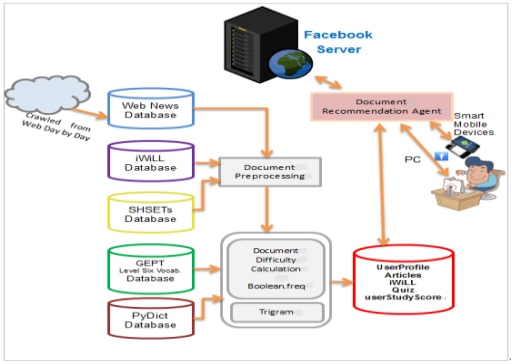


Fig. 1 The recommender system architecture diagram

### *A. System Login*

*1) Login with Facebook account*：The proposed system use *Facebook Open Graph API*, combined with a Facebook account to login system and authorize access user information. If user first login system, then get user information from Facebook to store in database and get user English degree from IWiLL database, otherwise, return user English degree from database.

*2) The recommended list generation method*：When the user clicks on articles menu or article category menu, the system randomly selects 5 articles or 5 articles of the same category that is appropriate for the user's English level for the user to read.

*3) Feedback mechanism*：Figure 2 is a scoring mechanism flow chart, indicating that learners rate the difficulty level after reading the article (1 to 6 points, 1 being the easiest and 6 is the hardest). This score is recorded and placed into the system to serve as the basis of the next recommended articles reference.

*4)* Users to interact with Facebook：The system combines Facebook Social Plugins components. When the user makes use of these components, such as commenting or liking, then the action message can also be synchronized to the user's Facebook wall.

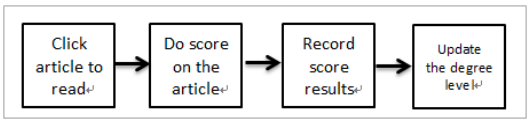


Fig. 2 Scoring mechanism flow chart

### *B. Web News collection*

In this study, the English news articles published by Taiwan News website are used as one of the most important sources for articles. The RSS Feeds service which have provided by these websites were used. Then a predetermined schedule executes a written program that grabs the latest news articles and content information daily from Taiwan News website to compare with articles in the database.

### *C. Data pre-processing*

*1)* Removal of superfluous words: For example, to grab and download from an English News Website articles, which contains HTML tags, such as <html>, </html> ... etc, and then to remove the HTML tags and non-English words.

*2)* Article cut sentence: English article letters convert to lowercase and then cut sentence.

*3)* Abbreviated word replacement: After article cut sentence, use replacement patterns to fix that by replacing contractions with their expanded forms, such as by replacing "can't" with "cannot", or "would've" with "would have".

*4)* Part-of-Speech Tagging(POS Tagging): Every word of the article labeled with its part of speech (for example: nouns, verbs, prepositions, ..., etc.).

*5)* Lemmatizing: Part of speech tag word restore to the original state (for example: plural noun revert to the singular noun).

*6)* N-gram：In this study, we use the Trigram method to cut a contiguous sequence of 3 words from a lemmatizing sentence before and after. Then we put phrases of three words

in the article are compared with PyDict dictionary to translate into Chinese, and store into the database, for a learner to learn the words or phrases.

### D. *Article difficult evaluation and grading*

Boolean.Freq is used to illustrate the diversity of vocabulary in each article. Using Boolean.Freq extended the value of weighted difficulty to calculate (GEPT 1 to 6), as shown in formula 1 and formula 2. Difficulty_1 is the difficulty level of the article including in the GEPT vocabulary set 1 to 6. Difficulty_2 is only the difficulty value of the GEPT 2 to 6 vocabularies. I is the level of the GEPT 1 to 6 vocabulary and it is also the weighted value of each level of the GEPT, and Gi is the number of GEPT 1 to 6 level vocabulary words in article.

$$Difficulty\_1 = \sum_{i=1}^{6} \frac{G_i * i}{G_i} \qquad \text{(Formula 1)}$$

$$Difficulty\_2 = \sum_{i=2}^{6} \frac{G_i * i}{G_i} \qquad \text{(Formula 2)}$$

By using Formula 1 for each article to calculate the degree of difficulty value (Difficulity_1), the degree of grade for each article is found, as seen in Table 1.

Table. 1 Articles grading

| difficulty_1 | degree |
|---|---|
| 0.0 ≦ difficulty_1 < 0.5 | 1 |
| 0.5 ≦ difficulty_1 < 1.0 | 2 |
| 1.0 ≦ difficulty_1 < 1.5 | 3 |
| 1.5 ≦ difficulty_1 < 2.0 | 4 |
| 2.0 ≦ difficulty_1 < 2.5 | 5 |
| 2.5 ≦ difficulty_1 | 6 |

## 3. Implementation

*1)* First link to the system home page (Figure 3) and click on the Facebook Login button using a Facebook account to logon.


Figure 3. System home page

*2)* Show the learners' homepage (Figure 4) after they have authorized access to provide an "activity feed" between friends and websites in Facebook for learners.


Figure 4. Learners' homepage

*3)* Figure 5 for the system to randomly select five articles suitable for the learners' English level to read or select the news category menu on the left.


Figure 5. Article recommended screenshot

*4)* Figure 6 is a Triple word or phrase that appears in the article for learners to study.

Figure 6. Triple word or phrase

*5)* Figure 7 sequence function is to join the fans group, click "Like" button on the article to agree, click "Recommend" button to recommend this system website they might like, sync to share with Facebook friends, and the article difficulty score.

Figure 7. like-box/button and ratings

*6)* Figure 8 is a comments window, similar to Message Board Function.


Figure 8. Comments window

*7)* Figure 9 is a query function of the learner's study history, where they can access the articles they have read previously


Figure 9. Learner's study history

*8)* Figure 10 is a query of the learner's friends on Facebook study history.


Figure 10. learner's friends on Facebook study history

9) Figure 11 shows how learners in the system can use the Facebook Social Plugins to post to Facebook. This function is synchronized to their Facebook wall and recommended to friends on Facebook, which can be treat as a "comment" function in a blog system or Facebook platform.


Figure 11. Learners' information on Facebook graffiti wall

## 4. Discussion and Conclusion

In the study, there were a total of 51,880 articles in the English news article database, a total of 149 students who participated, and a total of 200 articles were read, 152 articles have been ratings and feedback to the system. The system allows learners with a Facebook account can take advantage of Social Plugins components to interact with friends on Facebook by sharing articles and their opinions. Furthermore, some functionality have constructed in the proposed system, which can displaying the learning history of the user's Facebook friends, and see the articles they have read in the system. This allows the user to read more interesting articles and further their reading and learning ability. This structure has been the preliminary results and there are still need many areas that can be strengthened and improved. The following is the direction of future research: (1) CNN [12], BBC [13], China Post [14] and other well-known English news website article can be added in the future to provide learners with more diverse articles. (2) The future system can be turned into an application on Facebook, so that more users can search and use the system on Facebook. (3) Provide the function of searching for articles in the database with a specific keyword.

## Acknowledgement

This work described in this paper was supported by the grants from the National Science Council, Taiwan (Project No. NSC 99-2511-S-032 -004 -MY3)

## References

[1] Web 2.0, Available from http://zh.wikipedia.org/wiki/Web2.0.
[2] Facebook, Available from http://www.facebook.com.
[3] P. Resnick and H. R. Varian. Recommender systems. Communications of ACM 40(3), pp. 56-58. 1997. Available from http://doi.acm.org/10.1145/245108.245121.
[4] J. B. Schafer, J. A. Konstan and J. Riedl. E-commerce recommendation applications. Data Mining and Knowledge Discovery 5(1), pp. 115-153. 2001.
[5] S. D. Krashen, "The input hypothesis : issues and implications," 1985.
[6] The Input Hypothesis Model, Available from http://tw.myblog.yahoo.com/et-2005/article? Mid = 175& next=174 & l=f & fid=17.
[7] GEPT, Available from https://www.gept.org.tw/.
[8] Taiwan News, Available from http://www.taiwannews.com.tw.
[9] Daniel Gau, PyDict, English Chinese Dictionary, Available from http://sourceforge.net/projects/pydict
[10] IWiLL (Intelligent Web-based Interactive Language Learning ), Available from http://cube.iwillnow.org.
[11] NLTK (Natural Language Tolkit), Available from http://nltk.org/.
[12] CNN, Available from http://www.cnn.com.
[13] China Post, Available from http://www.chinapost.com.tw/.
[14] BBC, Available from http://www.bbc.co.uk/news/.