NEWordS: A News Search Engine for English Vocabulary Learning

Xuejing Huang, Sushma Chandra Reddy

Abstract: Vocabulary is the first hurdle for English learners to over- come. Instead of simply showing a word again and again, we come up with an idea to develop an English news article search engine based on users word-reciting record on Shanbay.com. It is designed for advanced English learners to find suitable reading materials. The search engine consists of Crawling Module, Document Normalizing module, Indexing Module, Querying Module and Interface Module. We propose three sorting & ranking algorithms for Querying Module. For the basic algorithm, five crucial principles are taken into consideration. Term frequency, inverse document frequency, familiarity degree and article freshness degree are factors in this algorithm. Then we think of a improved algorithm for the scene in which a user read multiple articles in the searching result list. Here we adopt a iterative & greedy method. The essential idea is to select English news articles one by one according to the query, meanwhile dynamically update the unfamiliarity of the words during each iterative step. Moreover, we develop an advanced algorithm to take article difficulty in to account. Interface Module is designed as a website, meanwhile some data visualization technologies (e.g. word cloud) are applied here. Furthermore, we conduct both applicability check and performance evaluation. Metrics such as searching time, word-covering ratio and minimum number of articles that completely cover all the queried vocabulary are randomly sampled and profoundly analyzed. The result shows that our search engine works very well with satisfying performance.

Keywords: Special Purpose Search Engine, Vocabulary Learning, News Retrieval, data mining

I. INTRODUCTION

A. Background & Motivation

For learners of the English Language, vocabulary is the first hurdle to overcome. To build up a solid foundation for the language, English learners keep reciting word list around the clock, which is quite a laborious job. Although it is an crucial step for ESL (English as a second language) learners to undergo, the so-called ”dictionary- memorization” is rather exhausting. Moreover, this method can inefficient to some extent if lack of sufficient and appropriate reviewing consolidation. For this reason, learners rack their brain for a better solution of vocabulary memorizing. Some students and English test taker keep flashcards or vocabulary notebooks. They go over these manuscripts periodically as reviewing. However, it seems that these kind of records are not fully made use of. Others turn to online English learning applications for help. One strategy adopted by some popular web applications (e.g. Shanbay) is to show a word again and again, which is somewhat mechanical. An- other method adopted by Baicizhan is to show an English word with a relevant picture. However, it is not so easy for some words to match a relevant picture. We can simply take the word ”ideology” as an instance. Under this kind of cir- cumstances, a picture with ambiguous meaning may distract users from normal word learning, or even misguide them to understand the word with a biased impression. There is al- so a technique which ask users to select the correct meaning of a word among many choices, but the meaning provided in a single choice is not adequate for comprehensively understanding a word. In reality, know all the meaning of a word will provide an overall view to a user, which is quite beneficial for vocabulary learning. In fact, users motivation of vocabulary learning is to both understand and use them in a real English context of reading, listening, writing and speaking. However the approach- es mentioned above are not capable to provide enough con- text for comprehension. It is apparent that reading itself helps. However, randomly chosen articles are likely to be too easy or too hard. Also, they may contain plenty of new words while few of the words which the learner is trying memorizing may appear in them. At this time, we associate the reading-related idea with the word-reciting record issues mentioned above. An effective way for learners to find suitable reading articles according to a digital word list accumulated by a user himself or herself will be very useful.

B. Relevant Work

In the last section, we have talked about features of popular vocabulary-reciting industrial products including Shan- bay, ToWords, Baicizhan and so on. Actually, few of them touch the field of reading. Though Shanbay launches a service called Shanbay Reading, the articles they provide to different users are completely the same. On the contrary, our idea is to develop an English news article search en- gine based on users online word-reciting record, in order to provide person customized service to individual users. Now we will turn to relevant academic works. So far, we have discovered some papers which have something to do with our topic. One category of them is specialized search engine including news search engine[5], which is a hot topic for a long time. Five related papers are adopted by ACM SIGIR 2015, which is being held right now in Chile. The other is about readingbased vocabulary expansion method.[1]. However, these are not identical with our idea at all. It is a news search engine for vocabulary learning that we struggle for. We believe that our project is pioneering in the field of information retrieval and will profoundly benefit English learners.

C. Objective and Goals

A news search engine is designed here to find suitable reading materials for advanced English learners. Based on modern information retrieval technology, we put forward the idea to build a specialized news search engine. It takes a list of word

- *Xuejing Huang, Computer Science and Technology, Stockholm University & Zhejiang University, Email: xuhu2343@student.su.se*
- *Sushma Chandra Reddy, Informatics under Multimedia Engineering, Stockholm University & Kaunas University of Technology, Email: sushcreddy@gmail.com*

as a query, meanwhile users retention degree is affiliated to every single word. It is enabled to returns selected and ranked results from authoritative news websites that can function as suitable reading materials for the user. Necessary information is presented at the same time. A user uploads his or her word-reciting record of Shanbay or vocabulary notebook to our search engine. Then the en- gine will screen on the Internet English news reports which are relative to the users word list. These news reports will be carefully selected and scored and will be eventually pro- vided by the engine to the user as excellently appropriate reading material according to his or her vocabulary note- book. Several specific goals are supposed to be achieved: A selected article should contains as many words in the query as possible under some constraints. Users can go through the whole article without too much comprehension and verbal obstacles. For articles with the same sum of word frequency, ones with more queried words are preferable. Factors shown below should be taken into considera- tion: the length and published time of an article, the grasping degree and the collection frequency of a word. Our search engine works quite better than simply sacking all the clusters of words into Google input box. We believe that our search engine does help.

D. Significance

Our search engine will help users find appropriate reading material in order to strengthen their grasp of vocabulary. Reciting a word again and again is basic way for language learners to expand their vocabulary, but they wont understand the meaning of words profoundly and will soon forget it without using it. Reading English news reports, which include the words a user is currently learning, will integrate new words into authentic English context. This is quite useful for vocabulary learning. Our search engine will pro- vide a ranked list of English news articles that are relevant to users vocabularyreciting record, and user can choose whichever from them. Reciting lifeless words again and again is quite a boring thing, which occasionally makes some learners to give up. By means of our search engine, however, a vocabulary learn- er can easily find interesting English news report according to their wordreciting record or vocabulary notebook. En- glish news reports are updated every day and have a lot of things to do with the real world. They are apparently more lively and attractive. Furthermore, provided with the chance to practice English words they have already recited, users will feel a sense of achievement and will carry on to struggle for more words.

II. METHODOLOGY

A. Prerequisite Requirements

Our aimed users vocabulary level should be above a designated threshold value. Since relevant research show that to read unsimplified material a learner of English needs to know at least 3,600 word forms (with a far higher number of meanings)[4], we do not recommend primers to adopt reading news as an approach to learn English actually. The engine receives a piece of word-reciting record or a vocabulary notebook as query. Each entry contains both a word and the degree of the users grasping of the word (or just a list of words, in which case the degree will be treated as the same). The number of entries in a single query varies from dozens to thousands. For the reason that the data used to be a query are recorded by software, the spelling of words should be correct.

B. Implementation Outline

This project consists of five main components: A crawler to download web pages from selective news websites, analyze and parse them. A document normalizer to do tokenizing and stem- ming. A indexing module to generate the inverted index. A querying module to evaluate how well document and query match and provide top K relevant documents as a ranked list. An interface module to receive search requests from the user and process them, return the results generated by the scoring module.

III. APPROACH SPECIFICATION

A. Crawling

This module is designed to fetch English news articles from the Internet. The crawler is implemented by means of Python 2.7. Urllib2, an extensible Python library for opening URLs is also employed. BeautifulSoup is also used here. In order to parse a document after page information being downloaded, XPath and Regular Expression are both used. Here in our prototype system, some authoritative news websites are selected as source, which are shown above. Washington Post: www.washingtonpost.com US News World Report: www.usnews.com The Guardian: www.theguardian.com New Scientist: www.newscientist.com The total of all crawled articles is up to 8,000.

B. Document Normalization

Crawled articles are normalized in this step. That is, we do some preprocessing upon the crawled English news arti- cles, which can be beneficial for the following steps. Tokenization is required here. At first we remove numbers and symbols, which will not appear in userss query (Shan- bay word-reciting record/vocabulary notebook). This will help compresses the article repository. Then, before we come to stemming, we try to do American- British word transforming. As is known to all, some English words are spelt with slight difference in America. For in- stance, colour is the British form of the English word, but it can be spelt as color in American English. Such kinds of distinguishments between British and American English is somewhat annoying for further stemming processing, meanwhile seldom touched in papers and projects. So here we develop a Python program to deal with distin- guishments between British and American English. British forms are simplified to its corresponding American ones. The pairs of British-American transformation that are con- sidered in our work are shown in Table 1 After that, we do stemming using the Lovins stemming algorithm[2]. Compared with the original Porter stemming algorithm, the Lovins algorithm is much faster. It has ef- fectively traded space for time, and with its large suffix set it needs just two major steps to remove a suffix, compared with the eight of the Porter algorithm.

C. Indexing

It is obviously unrealistic to do ad-hoc searching. So we do indexing in advance to organize document and word information. First, we store key information of the downloaded articles in a linked list. Each node of the list includes ID, URL, published date and source of the article. After that, Inverted index will be built for the words that appears in the documents.

For each term t, we store a list of all documents that contain t. Considering that our En- glish news search engine is for vocabulary learning and has nothing to do with semantic meaning of the articles, we in- troduce Bag-of-words Model (BOW) here. In this model, a document is represented as the bag (multiset) of its word- s, disregarding grammar and even word order but keeping multiplicity. Then, we use a hash table rather than a trie to store the dictionary in index. The main advantage of hash dictionary is its fast speed. Confliction will be avoided and time com- plexity of hash lookup is O(1). Since stemming has been done, there is no need to find similar words. In the mean- time, vocabulary here is limited, we dont have to rehash the table. So, we can overlook the shortcomings of hashing. Furthermore, since only disjunctive search is possible to appear in the applications scenarios of our projects, skip pointers are not adopted.

D. Querying

This is the most essential component of our project. It is aimed to finding the most suitable articles according to the query. Result will be sorted based on their score. Technical specification will be delivered here.

REFERENCES

[1] J. R. Joseph, M. L. Stein, and K. Wysocki Learning vocabulary through reading, American Educational Research Journal, 21(4):795825, Winter 1984.

[2] J. B. Lovins. Development of a stemming algorithm. Mechanical Translation Computational Linguistics, 11:2231, 1968.

[3] C. D. Manning, P. Raghavan, H. Schu tze, et al. Introduction to information retrieval, volume 1. Cambridge university press Cambridge, 2008.

[4] T. Saragi. Vocabulary learning and reading. System, 6(2):7278, 1978.

[5] B. Zhang, Z. Zhao, L. Zhang, and L. Weng. Building a specialized search engine of special subject. TENCON 02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 1:6972, 2002.