A Cross-Domain Recommender System Based on Common-Sense Knowledge Bases

Yi-Ting Tsai*, Chih-Shiang Wu†, Hsiang-Ling Hsu†, Tenniel Liu†, Pei-Lin Chen†, Wen-Hao Chen* Keng-Te Liao*
*Department of Computer Science and Information Engineering
National Taiwan University
{aliciatsai, b02902023, d05922001 }@ntu.edu.tw
†Industrial Technology Research Institute
{itri531050, sharonhsu, tenniel.liu, mia-chen}@itri.org.tw

***Abstract*—A system able to extract and recommend technical terms from various domains is proposed in this paper. The motivation is to provide keywords that users may not be familiar with in the beginning but will be interested in after studying. To acquire domain knowledge, we collect documents from various sources, and the words in the documents are then represented as semantic word vectors. Given queries from users, the system first extracts important terms from given documents and computes the semantic similarity between those terms. Next, we utilize third party common-sense knowledge bases such as ConceptNet and Wikipedia to connect the queries to those extracted keywords through the network structures. Finally, the system will collect all keywords traversed and recommend the top-n of them.**

We propose and compare four models for the recommendation, and the differences between using ConceptNet and Wikipedia for discovering related knowledge are also investigated in this work.

***Index Terms*—sense knowledge, recommender, information extraction**

I. Introduction

When engaging in creative thinking, people tend to link different ideas from various domains. Nevertheless, connecting ideas from different domains is rather hard. To effectively connect ideas from different domains, one needs to possess enough professional knowledge within the field. Therefore, this paper aims at designing and implementing a recommender system that extracts useful and critical information from documents in different domains and recommends "related yet cross-domain" keywords from other fields using Chinese common-sense knowledge base. Using unsupervised machine learning techniques, the system will expand the user's query by computing its similarity with words in common-sense knowledge base. Intermediary terms from query expansion are later used as inputs to search for cross-domain related keywords that the system will recommend for the user.

Unsupervised information extraction techniques largely use statistics and linguistic methods to compute the importance of each word. For instance, to identify important keywords, we may look for technical terminologies or terms that appear frequently in a document [1], [2], or use grammatical analysis [3] or sentence clustering [4], [5]. One advantage of these methods is that there is no need to obtain human-labeled training data. The above methods can also be used to train data from different domains and different languages. However, the disadvantage is that they rely too much on the occurrence of a word. If a keyword rarely appears in the documents, it is likely to be overlooked. Furthermore, statistical methods are suitable for long documents but not for short documents. They usually perform unsatisfactory on the short documents, where statistical information is inadequate.

We choose the supervised information extraction method as it provides better recommendation performance and we try to overcome the disadvantage by designing an "automatic labeling" module for the system to reduce the labeling time.

In this paper, we propose a recommender system that combines common-sense knowledge base with a search engine built upon word embedding. Users can search by simple query to get a more comprehensive multi-domains results. The system uses word embedding to link information extracted from different domains. Thus, after query expansion, the system can more accurately recommend cross-domain results to the users.

After reviewing related works and systems in Section 2, we continue to introduce the proposed architecture for our cross-domain recommender system in Section 3. Then in Section 4, we implement the proposed system and conduct corresponding experiments in Section 5. At last, we discuss the results and conclude our current work and future work.

II. Related Work

For Chinese common-sense knowledge base, there are "Chinese WordNet" [1], "E-HowNet" [2], "ConceptNet" from MIT Media Lab [6], [7] and "Traditional Chinese Wikipedia" from Wikipedia.

Related Chinese-based knowledge recommender systems include "RM i5E Platform" [3] developed by Industrial Technology Research Institute(ITRI). Another work from Yuan, et al (2015) [8] uses ConceptNet common-sense knowledge base to assist experienced or inchoate designers to improve the quality of reframing and frame creation process. In the paper of Cambria, E. et al (2010) [9], the author combines ConceptNet and WordNet to form a new "semantic network".

[1]Chinese Wordnet: http://lope.linguistics.ntu.edu.tw/cwn2/
[2]E-HowNet: http://ehownet.iis.sinica.eud.tw/
[3]RM i5E Platform: http://www.ithome.com.tw/people/108339

The author later uses this grouping result to analyze sentiment of a given document.

III. Cross-Domain Recommendation Architecture

Firstly, we need to obtain "sematic network" from both third party common-sense knowledge base and our own labeled knowledge base. Our own labeled knowledge base came from documents provided by Industrial Technology Research Institute; therefore, we will refer it as the ITRI knowledge base. The system has an "automatic labeling" module labeling keywords from the given documents. We will use the labeled keywords to train our own knowledge base.

The system would be designed to recommend related technical keywords from our own knowledge base based on a user's query. To accomplish this goal, the system has a "cross-domain recommendation" module that firstly expands user's queries using several proposed query expansion methods. Next the system will use those expanded intermediary terms to search for related technical keywords in our own knowledge base.

Here, we propose four recommending methods. The first two skip query expansion and recommend keywords directly from the knowledge base. The next two methods firstly perform query expansion and then use expanded intermediary terms to search for keywords in the knowledge base.

Figure 1 and figure 2 illustrate the structures of our cross-domain recommendation system. We name the proposed four methods as M1, M2, M3, and M4. M1 will recommend related technical keywords directly from the ITRI knowledge base. M2 will recommend related keywords directly from ConceptNet. M3 and M4 will perform query expansion using ConceptNet, and Wikipeida first. And we will obtain related intermediary terms form the above two common-sense knowledge bases. The system will then use these intermediary terms to search for related technical keywords in the ITRI knowledge base.

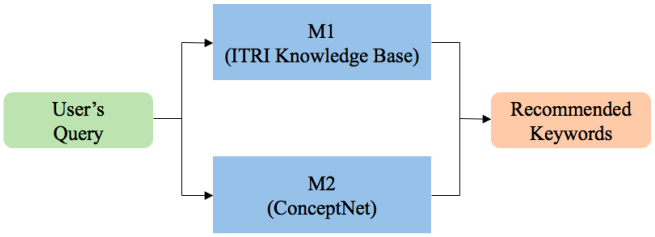

Fig. 1. cross-domain recommendation architecture (M1 and M2)

IV. Implementation of the Recommender System

We will firstly address how to construct our own ITRI knowledge base. Next, we will give detailed explanation about how to construct the "cross-domain recommendation" module using word embedding.

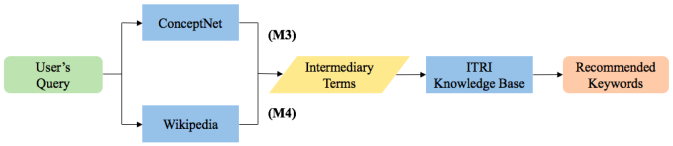

Fig. 2. cross-domain recommendation architecture (M3 and M4)

A. ITRI Knowledge Base

To train our own knowledge base, we need to preprocess our documents to obtain a corpus from the given documents. Documents provided by ITRI have two main categories, documents with technical information and with general needs information. Technical documents are obtained from the following sources: MaterialNet [4], INSIDE [5], TechOrange [6] and documents from ITRI. General needs documents are obtained from the following sources: MyDesy [7], CommonHealth Magazine [8], Parenting Magazine [9], Global Views Monthly Magazine [10], and Ministry of Health and Welfare [11].

Figure 3 illustrates the process of constructing the ITRI knowledge base. After obtaining the documents, we use the CKIP (Chinese Knowledge and Information Processing) word segmentation service to perform word segmentation and obtain our training corpus that includes both technical keywords and general needs keywords. Finally, we can construct the ITRI knowledge base from the obtained corpus using the word2vec toolkit [10]. The knowledge base will then have both general needs and technical keywords that are ready for the recommender system.

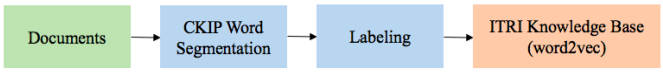

Fig. 3. Process of Constructing ITRI Knowledge Base

B. Cross-domain Recommendation

To combine the common-sense knowledge base and our own ITRI knowledge base, we use four corpora from Wikipedia, ConceptNet, ITRI technical documents, and ITRI general needs documents to train their word embedding using the word2vec toolkit.

Using word embedding enables us to quantify the similarity between each word so that we can combine various semantic networks. In particular, the similarity is measured by cosine similarity score.

[4]MaterialNet: https://www.materialsnet.com.tw/
[5]INSIDE: https://www.inside.com.tw/
[6]TechOrange: https://buzzorange.com/techorange/
[7]MyDesy: https://www.mydesy.com/
[8]CommonHealth Magazine: http://www.commonhealth.com.tw/
[9]Parenting Magazine: https://www.parenting.com.tw/
[10]Global Views Monthly Magazine: https://www.gvm.com.tw/
[11]Ministry of Health and Welfare: https://www.mohw.gov.tw/mp-1.html

The use of the four corpora can provide the system with various benefits. We use Wikipedia and ConceptNet to expand the user's original query and obtain more related words. It is expected that more intermediary terms related to the needs keywords in the ITRI knowledge base so that the system can search for more diverse keywords using intermediary terms. By training needs keywords together with the technical keywords in the same knowledge base, we are able to search for related technical keywords whose average similarity to those expanded intermediary terms is higher. Table I contains basic information of the four corpora introduced above.

TABLE I
BASIC INFORMATION FOR EACH CORPUS

Information	Corpus			
	Wikipedia	***ConceptNet***	***ITRI needs documents***	***ITRI technical documents***
Word Embedding Algorithm	word2vec CBOW	ConceptNet Numberbatch [7]	word2vec CBOW	word2vec CBOW
Number of Chinese Words	660,000	50,000	36,000	48,000
Keywords for Training Model	25,103,000	5,513,000	3,045,000	2,622,000

C. Implementation Process for the Four Methods

Figure 4 and figure 5 summarize the implementation process for M1 and M2.

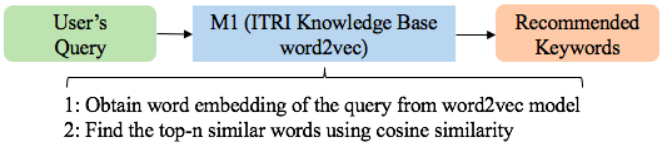

Fig. 4. Implementation process for M1

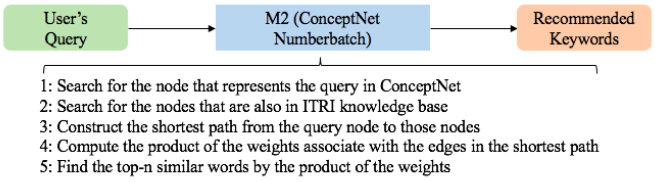

Fig. 5. Implementation process for M2

Next, we will discuss how to obtain the final search results using the expanded intermediary terms from ConceptNet, and Wikipeida.

After obtaining the expanded intermediary terms, we will use them as input queries to search for final related technical keywords in the ITRI knowledge base. But, how can we measure the similarity between each word in the ITRI knowledge base with all the related intermediary terms? We can easily get the final top-n related technical keywords by sorting all the similarity score in the word vectors. Nevertheless, this method may result in a significant problem that the final results may only be related to specific one or two intermediary terms rather than all the intermediary terms. Then this will lose the original purpose of expanding user's query using commonsense knowledge base. Therefore, to avoid the problem, we use average cosine similarity to determine the final top-n results.

Figure 6 illustrates the implementation process of M3 and M4, and figure 7 shows how to calculate the average cosine similarity. $Q_1, Q_2, ..., Q_m$ denotes the intermediary terms. $W_1, W_2, ..., W_n$ represents all the words in the ITRI knowledge base. The similarity score between each pair of Q's and W's word vectors is calculated and averaged in column direction. After the calculation, $W_1, W_2, ..., W_n$ are sorted according to the scores, and are denoted as $X_1, X_2, ..., X_n$. We then choose the final top-n keywords to recommend from $X_1, X_2, ..., X_n$. The averaging method ensures that the recommended words are related to all intermediary terms rather than only a few of them.

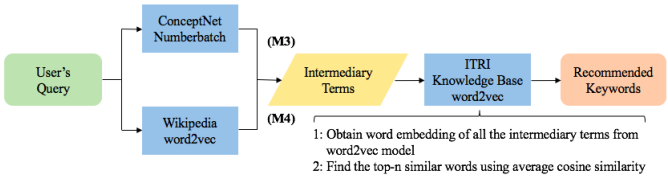

Fig. 6. Implementation process for M3 and M4

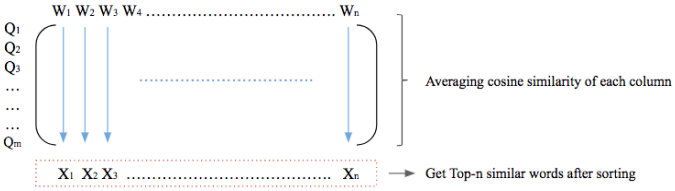

Fig. 7. Calculating Average Similarity

V. EXPERIMENT

We select one testing query to show the results of our proposed recommender system. For M1 and M2, the system will output the final 10 keywords directly. For M3 and M4, the system will first obtain 10 related intermediary terms through query expansion, and use these 10 intermediary terms to search for the final 10 keywords. The final 10 keywords are then recommended to users.

Tabel II shows the results of the testing query, 藥物(medicine). Since the system aims at building a Traditional Chinese based recommender system, the results will be shown in Traditional Chinese coupled with English translation.

For more results, please see the link or our demo website.

TABLE II
RESULTS USING TESTING QUERY 藥物(MEDICINE)

Methods	Return keywords
M1 (ITRI knowledge base)	水楊酸(Salicylic Acid) 蓖麻油(Caster Oil) 光波導(Optical Waveguide) 氧化(Oxidation) VitriBand (Cell-Free Bandage-Type Artificial Skin) PET (Positron Emission Tomography) 微藻(Microalgae) DEHP (Bis(2-ethylhexyl) phthalate) 人工皮(Artificial Skin) 甲醇(Methanol)
M2 (ConceptNet)	藥物(Medicine) 藥劑(Pharmaceutics)
M3 (ConceptNet + ITRI knowledge base)	**Intermediary terms:** same as M2
	Recommended keywords:
	水楊酸(Salicylic Acid) 蓖麻油(Caster Oil) 人工(Artificial) DEHP (Bis(2-ethylhexyl) phthalate) 甲醇(Methanol) 光波導(Optical Waveguide) 氧化(Oxidation) 尿素(Urea) PET (Positron Emission Tomography) VitriBand (Cell-Free Bandage-Type Artificial Skin)
M4 (Wikipedia + ITRI knowledge base)	**Intermediary terms:**
	抗生素(Antibiotic) 用藥(Medication) 藥品(Drug) 類藥物(Drugs) 鎮靜劑(Sedative) 處方(Prescription) 抗病毒(Antiviral) 療法(Therapy) 該藥(Medicine) 藥劑(Pharmacy)
	Recommended keywords:
	水楊酸(Salicylic Acid) 人工皮(Artificial Skin) 蓖麻油(Caster Oil) 甲醇(Methanol) 尿素(Urea) DEHP (Bis(2-ethylhexyl) phthalate) VitriBand (Cell-Free Bandage-Type Artificial Skin) PET (Positron Emission Tomography) DBP Oxidation (Degradation of Dibutyl Phthalate) TDI Phthalates (Phthalates Tolerable Daily Intake)

VI. Discussion and Conclusion

The final results from the above experiments show that the system serves its purpose of recommending related keywords for the users. However, the recommended keywords will be slightly different depending on the methods.

M1 gives keywords ranging from general keywords to more technical keywords. M2, however, gives only very general keywords since ConceptNet knowledge base does not contain many technical keywords.

M3 and M4 expand the queries using ConceptNet and Wikipedia. These will provide 10 intermediary terms that are related to our original queries. The system will recommend 10 final keywords with highest average cosine similarity to all the 10 intermediary terms. After expansion, we can see that the system recommends keywords slightly differently. M3 generally produces keywords that are less technical than M4 since intermediary terms obtained from ConceptNet are less technical than Wikipedia.

As can be observed from the results of the intermediary terms of Table II and Table III, ConceptNet usually provides more general vocabularies yet Wikipedia outputs more technical terms. Since Wikipedia contains more technical and professional details on the web page, the intermediary terms obtained from Wikipedia can result in a better search for the technical keywords from the ITRI knowledge base.

A final note is that the ITRI knowledge base is rather small compared to ConceptNet and Wikipedia. This could explain why even though the system is able to recommend differently based on different methods, the recommended keywords are greatly overlapping. In the future, we plan to generate more labeled data with the "automatic labeling" module. We believe that the system will have a better performance with more labeled data. Besides, we will also explore the possibility to combine or modify the four proposed models to improve the quality of recommendation.

VII. Acknowledgment

The authors would like to thank the reviewers for their helpful comments. The authors would also like to thank Taiwan's Ministry of Economic Affairs for financial support, ITRI for resource support and all the participants in Dechnology team from ITRI for their help in making this work possible.

References

[1] HaCohen-Kerner, Yaakov. "Automatic extraction of keywords from abstracts." International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2003.
[2] Uzun, Yasin. "Keyword extraction using naive bayes." Bilkent University, Department of Computer Science, Turkey, 2005
[3] Pasquier, Claude. "Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation." In Proceedings of the 5th international workshop on semantic evaluation. Association for Computational Linguistics, 2010.
[4] Kuo Zhang, Hui Xu, Jie TangJuanzi Li "Keyword extraction using support vector machine."International Conference on Web-Age Information Management. Springer Berlin Heidelberg, 2006.
[5] Zhang, Chengzhi. "Automatic keyword extraction from documents using conditional random fields." Journal of Computational Information Systems 4.3 (2008): 1169-1180.
[6] Robert Speer and Catherine Havasi. "Representing General Relational Knowledge in ConceptNet 5." LREC, 2012.
[7] Robert Speer, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." AAAI Conference on Artificial Intelligence, 2017.
[8] Yuan, Soe-Tsyr Daphne, and Pei-Kang Hsieh. "Using association reasoning tool to achieve semantic reframing of service design insight discovery." Design Studies 40 (2015): 143-175.
[9] Erik Cambria, Robert Speer, Catherine Havasi, Amir Hussain "SenticNet: A Publicly Available Semantic Resource for Opinion Mining." Artificial Intelligence, 14–18 (2010).
[10] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean "Efficient Estimation of Word Representations in Vector Space", In Proceedings of Workshop at ICLR 2013