MindMover: Associating Chinese Concepts for Cross-Domain Recommendation

Zheng Wang, Kailang Jiang, Kenny Q. Zhu

Shanghai Jiao Tong University, Shanghai, China
wangzhdemail@gmail.com,cheerjiang510@gmail.com,kzhu@cs.sjtu.edu.cn

Abstract. This paper demonstrates a Chinese association knowledge-base which allows the machine to associate one concept or object to another, just like how imagination runs in a human mind. For example, it is able to answer the question, "What do you have in mind when you think about 'Gong Li'(巩俐)." One important application of the association knowledge-base is cross-domain recommendation. This association knowledge-base can significantly improves the diversity in recommendation and discover latent connections between items so as to infer the profile of the users and predict future needs.

1 Introduction

Knowledge is the high level, structured abstraction from basic facts. Recently, there has been many attempts to extract and construct knowledge from unstructured text data[7, 4]. In this paper, we present a Chinese association knowledge-base, which is extracted from three Chinese encyclopedias, Chinese Wikipedia [3], Baidu Baike [1], Hudong Baike [2]. The articles in these three encyclopedias are used to discover associations between two terms, such as "party" and "beer." The relationship between these two terms can be explained as a conditional probability $p(party|beer)$, or the likelihood that one thinks of "party" when mentioning "beer." The probability is derived generally from the sentence-level co-occurrence within the three corpora and the co-occurrences between the title and the body of the encyclopedia articles. We treat these two kinds of co-occurrences separately because we believe they carry different aspects of the association. The whole knowledge base can be viewed as a weighted directed network where Chinese terms are nodes and directed edges are the associations.

This comprehensive, open-domain association network can be used in many scenarios [5, 6]. One of them is *cross-domain* recommendations in e-commerce. User A can be represented by a list of commodities or services he or she has purchased, i.e., $t_1, \ldots, t_n$. We can then carry out two kinds of recommendation. For content-based recommendation, we can calculate the likelihood of the user to purchase the next product t_{n+1} as $p(t_{n+1}|t_1, \ldots, t_n)$, which maybe approximated by $p(t_{n+1}|t_1)$, $p(t_{n+1}, t_2)$, etc. For collaborative filtering recommendation, we can calculate the resemblance between user A and user B by connecting the items purchased by A and B through the association network.

Next we introduce the three corpora, then outline the construction of the knowledge base, and finally explain the demo setup.

2 Three Chinese Encyclopedias

The data sources for our association knowledge-base are Chinese Wikipedia, Baidu Baike and Hudong Baike. Each article in three encyclopedias is an unit for processing. The number of articles in Chinese Wikipedia, Baidu Baike, Hudong Baike are 1 million, 7 million, and 4 million respectively. All articles in data source are contributed by community authors. Each article has a title which is a concept or an entity and a body text which is used to describe the title.

The articles in three encyclopedias may have the same title. For example, three encyclopedias all have the article about "Shanghai." But these three articles are different in the body text. We first count those two kinds of co-occurrence in each encyclopedia separately, and then merge each kind of co-occurrence in three encyclopedias together. Then we get the merged sentence-level co-occurrence and merged title-body co-occurrence. Besides, there are synonyms in our dataset. We linked the synonym to a synset. The synset has a representative title and a list of terms which are the synonyms of the representative title. So we also merge co-occurrence related to words in the same synset together. Finally, we get the merged sentence-level and title-body co-occurrence between synsets.

3 The Association Knowledge Base

Our input dataset consists of 8 million of synsets. We first compute the association relationship between any two synsets and then generate association knowledge-base.

The association between two concepts or synsets in association knowledge-base is computed from two kinds of co-occurrence. The first is sentence-level co-occurrence, which means two terms co-occur in a sentence in any of our documents. The second is known as title-body co-occurrence, which means one term appears in another term's description page. Given two synsets A and B, the probability of $P(B|A)$ is defined as follows:

$$P(B|A) = \frac{CoOccur_{s\text{-}level}(A,B) + CoOccur_A(B)}{\sum_{i=1}^{n} CoOccur_{s\text{-}level}(A,S_i) + length(A)} \tag{1}$$

Here $CoOccur_A(B)$ represents the frequency of B appearing on A's page, and $length(A)$ means the total length of A's page, while S_i can be any synset that co-occurs with A in a sentence.

For each synset, we can get a list of related synsets ordered by the probability score described above, and those related synsets can usually be of great variety. While sometimes we need a special kind of synsets from the results to achieve a special goal, so we do clustering on those synsets. And since each term

has several category labels on its encyclopedia page, we collect those category information, cluster synsets that have common category labels together, and represent each cluster with those labels. If cluster A has category labels c_1 and c_2, while cluster B has label c_3, and synset s has labels c_1,c_2,c_3 andc_4, then we compute a relatedness score between s and A,B as $R(s, A) = P(c_1|s) + P(c_2|s)$, and $R(s, B) = P(c_3|s)$, and assign s to the cluster with a higher relatedness score with s. After clustering, we can get each cluster which is also an ordered synsets list by its category labels.

4 System Demo

The demonstration consists of two parts. The first showcases the association knowledge base. The second illustrates the application of the knowledge base in a recommendation system.

Figure 1 demonstrates the result of querying the association knowledge base. The famous Chinese actress "Gong Li" is used as a query concept. And all related concept for "Gong Li" are depicted in a clustered manner. When you type a concept into our system, the system will return related concepts which is exhibited in a clustered manner. The closer the relationship, the shorter the distance between the two concept in the figure. The more popular the related concept, the bigger the font used to display this concept.

Fig. 1. Concepts Associated with Actress "Gong Li" and Their Clusters.

In this part of the demo, we showcase a prototype recommendation system which uses both content-based recommendation and the more popular collaborative filtering. In content-based recommendation system, after the system generates a user model from inputs (purchase history), association knowledge helps

to expand the user model and generate novel recommendations. In collaborative filtering case, the knowledge base helps the system relate two users who have latent connection or similarity.

Fig. 2. Association Based Collaborative Filtering

Figure 2 illustrates the use of association in collaborative filtering recommendation system. At the first glance, $userA$ and $userB$ are unrelated because they have different purchase histories. When applying the association knowledge base, the latent relationship between $userA$ and $userB$ could be discovered. The CD and car cleaner are all related to the concepts "car" and "trip". The tent and beer indicates a trip to camping, which in turn, also related to "car." Thus the latent relationship between $userA$ and $userB$ is discovered. Then the system recommends the CD to $userB$ and recommends a tent to $userA$.

References

1. Baidu encyclopedia. `http://baike.baidu.com/`.
2. Hudong encyclopedia. `http://www.baike.com`.
3. Wikipedia (chinese). `http://zh.wikipedia.org`.
4. J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, 194:28–61, 2013.
5. S. Jiang, D. Lowd, and D. Dou. Learning to refine an automatically extracted knowledge base using markov logic. In *ICDM*, pages 912–917, 2012.
6. J. Tang, S. Wu, J. Sun, and H. Su. Cross-domain collaboration recommendation. In *KDD*, pages 1285–1293, 2012.
7. W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD Conference*, pages 481–492, 2012.