The hitchhiker's guide to recommendations

Introducing a recommendation system to an online collaborative dictionary

Eser Aygün
Computer Engineering Department
Istanbul Technical University
Istanbul, Turkey
eser.aygun@itu.edu.tr

Amaç Herdağdelen
Center for Mind and Brain Sciences
University of Trento
Rovereto, Italy
amac.herdagdelen@unitn.it

ABSTRACT

We studied a very popular online hyper-textual collaborative dictionary in Turkish called İTÜ SÖZLÜK. Previous studies show that one of the properties of such knowledge bases created by a large number of people is the power-law distribution of the in-degrees of the topics. Many topics are created daily and they are forgotten the next day without receiving any hyper-references. However, existing links created by the users may help us in determining related topics or evaluating other recommendation systems. We crawled İTÜ SÖZLÜK, lemmatized the words using a morphological analyzer and constructed a bag-of-words based vector space to calculate relatedness of topic pairs. The relatedness scores are used to populate recommendations for topics as *other topics that may be related*. Preliminary results suggest that by using content-based models, we can boost the average number of recommendations per title at least by a factor of three without compromising the recommendation quality when compared to a currently used graph-based model.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

algorithm, experiment, performance

Keywords

data mining, itü sözlük, online dictionary, recommendation, Turkish

1. INTRODUCTION

İTÜ SÖZLÜK[1] is a rapidly growing online, hyper-textual, collaborative dictionary in Turkish. Its philosophy is inspired by Douglas Adams' *The Hitchiker's Guide to the Galaxy* [1]. In Turkey, it is a very popular web site with an Alexa rating of 4375 as of May 2009. Everyday, more than 6,000 registered users are creating thousands of entries about quite diverse topics –*titles* as they are called in the site's jargon. The users are allowed to include hyper-references in their entries which links a title to another indicating a semantic, contextual, phonological or any other relevant association between the titles. There are no constraints on the content of titles or entries as long as they contain legal materials and abide the dictionary format. Since the day the site was opened (March 2004) until when this manuscript was written, more than two million entries grouped under nearly 400,000 titles were written.

If we consider the titles as the nodes of a graph and the hyper-references as directed links between these nodes, we obtain a large, evolving network of titles. Previous work indicates that this kind of socially constructed topical networks exhibit scale-free properties [6, 2]. To give an impression of the heavy tail of the degree distribution, in the snapshot of the dataset that we worked on, there are more than 140.000 titles (almost half of the titles) with fewer than or equal to 5 incoming links. This poses a problem for a site one of whose major goals is to be an unconventional information source on practically anything the users want to talk about: Titles should be accessible by following links created by those who are interested in similar titles.

To solve this issue, one may employ a recommender system, such as one which shows a list of "possibly related" titles to the one currently being viewed by the user. Then the problem reduces to, given a title, finding relevant titles that would attract the attention of the users looking at given title. Thus, recommendations would act like artificially created hyper-references. İTÜ SÖZLÜK, recently employed such an approach by recommending the users five other titles which have a hyper-reference pointing to the currently viewed one. Although no systematic evaluation is done, the owner of the site reports a "substantial increase in number of clicks" when using the backward links (i.e. the links that point to the title)[2].

However, such a system has several drawbacks. If a title has no incoming links, then no recommendation can be done for it and this is the case for a majority of the titles, especially the recently created ones. The problem of finding recommendations for items in the heavy tail and cold start (i.e. recently arrived/created items) are known problems in

[1]http://www.itusozluk.com/

[2]Private communication with Çağatay Gürtürk, the owner of İTÜ SÖZLÜK.

RecSys '09 New York, NY USA

recommender systems literature albeit for different domains [9, 5].

The contribution of this study is two-fold. First, we implement and compare two vector spaces to represent titles in a high dimensional space. Vector spaces are a widely used technique to represent concepts, words, or documents in computational linguistics and information retrieval [10, 8, 7]. In our case, vector spaces also help us to overcome the sparseness in graph structure and they augment the coverage of the recommendations. The first vector space is a classical bag-of-words model where the dimensions are the words found under a title. The second vector space is an implementation of the Explicit Semantic Analysis model [3]. The performance of these content-based models are tested according to their predictive power on the existence of a hyper-reference between a pair of titles.

The second contribution is that we also carried out a small-scale experiment on human subjects (a sample of the users of the site) to judge the relatedness scores between title pairs. By using these ratings, we were able to compare the content-based models with the models depending on the link structure of the nodes.

An additional contribution is the introduction of a novel dataset in Turkish. To our knowledge, only a few recommendation systems evaluation were done for Turkish language [4]. In this study, we employed lightweight linguistic processing (only a Turkish lemmatizer is used), therefore it is straightforward to implement the models we study in similar datasets in other languages.

2. METHODOLOGY

Our methodology can be separated into three steps. First, we crawled and processed İTÜ SÖZLÜK. Then, we proposed different recommendation models and, finally, we evaluated proposed models using two different evaluation techniques. Following subsections give further details about these three steps.

2.1 Data

The basic unit of content in İTÜ SÖZLÜK is an entry written by a user. The entries are organized under titles and listed chronologically from older to newer. There may be more than one entry under a title, while each entry is associated with exactly one title. We crawled a snapshot of İTÜ SÖZLÜK as of June 2008. In that snapshot, there were 1,415,638 entries organized under 305,399 titles. By parsing the hyper-references, we extracted the graph of titles containing 1,271,239 directed links.

Each title, being a node of the graph, can also be interpreted as a text document generated by concatenating its entries. Thus, we were able to apply well-known document classification techniques on İTÜ SÖZLÜK titles. First, we lemmatized every word in the corpus using Zemberek[3], a Turkish natural language processing package. This process led to 201,978 distinct unigrams, which were further filtered by their frequencies to determine the terms. The terms were selected within the unigrams by filtering out the most frequent 100 and storing the next 20,000. All titles were then represented by a vector with a dimensionality of 20,000, whose components correspond to the term frequencies –the number of occurrences of the respective term inside the respective title. We called these vectors *term vectors*.

To be able to adapt the approach introduced in [3], we also determined the *concepts* that are assumed to represent the conceptual space of the İTÜ SÖZLÜK content uniformly. The concepts were chosen to be the first 20,000 titles with the highest in-degrees that contain at least 100 unigrams. In order to focus our evaluation on the titles with few links we constructed another set of titles called the *targets*. The targets are the titles that are subject to recommendations. In other words, recommendations are constrained in the domain of targets. We decided to limit the in-degrees of targets to 5 to make sure we are dealing with the heavy tail of the degree distribution. Therefore, the targets were chosen to be the first 10,000 titles with the highest number of unigrams which have not more than 5 incoming links.

In the final phase of the data preparation step, we weighted the terms using the term frequency - inverse document frequency (TF-IDF) measure [11], which considers both term frequency and term distinctiveness at the same time.

2.2 Models

In our study, we formulated the problem of recommending titles as identifying related titles. To bring out the relatedness of title pairs, we built four different models relying on either graph-based or content-based properties of the dictionary. There is also a fifth model, namely *the random model* (*RND*), which we use for random-performance baseline.

The inputs of graph-based models are limited to the title graph, whose nodes and links consist of titles and hyper-references, respectively. The first such model is the the one that is actually being used by İTÜ SÖZLÜK and we called it *the back-link model*, or *BCK*. Given a target title t, BCK recommends all titles that refer to t. BCK does not generate scalar relatedness scores between targets, so it cannot recommend arbitrary number of titles nor it can sort its recommendations by their importance. These problems prevent us from using BCK effectively.

To tackle the problems of BCK, one can use the number common referrers and referees of two titles as a measure of relatedness between them. A common-referrer (common-referee) of a pair of titles is a title which has a hyper-reference to (is referenced by) both of the titles. Our experiments revealed that using the number of common referrers for initial scorings and then breaking the ties by the number of common referees worked best. We called this combined model *the referral-based model*, or *REF*.

Graph-based models disregard a large portion of the available data, namely the content. Content-based models, on the other hand, make use of the texts entered under the titles to measure the relatedness scores. We adapted two of the state-of-the-art document classification methods into our problem. The first content-based model is founded on the well-known bag-of-words approach. In this model, the relatedness scores are measured simply by calculating the pairwise cosine similarities between the weighted term vectors of the targets. We called this model *the term-based model* (*TRM*) as it works on the term space.

The second content-based model is an adaptation of the work of Gabrilovich and Markovitch[3]. According to Gabrilovich and Markovitch, the vectors representing the documents can be transported from the term space into a concept space to reach more explicit representations. They de-

[3]http://code.google.com/p/zemberek/

fine the concepts as "the basic units of meaning that serve humans to organize and share their knowledge"[3] and use WIKIPEDIA[4] articles as a source of such "basic units of meaning". By processing WIKIPEDIA articles corresponding to the concepts, they build and store a concept-term matrix $\mathbf{C}$. Then, given a term vector $\mathbf{p}$ representing a document d, they compute $\mathbf{q} = \mathbf{p} \times \mathbf{C}^T$, which is the concept vector of d. The document-document comparisons are carried over those concept vectors rather than term vectors. We applied the same approach considering the titles as documents, where the content of a title is the concatenation of the entries written under it. In our application, however, there was not an external source for concepts but İTÜ SÖZLÜK itself. Our concepts were the first 20,000 titles with the highest in-degrees containing at least 100 unigrams. The rationale for this decision is that a high number of incoming links for a node means that the corresponding title is highly referenced in other contexts which is an indication of being conceptually fundamental.

2.3 Evaluation

We evaluated our models with two experiments. The first experiment focuses on to what extent the content-based models' relatedness score align with the current link structure of the graph. In the second experiment, we populate a set of recommended pairs by using each model, collect human judges' relatedness scores for those pairs, and contrast the mean score for each model. The details are given below.

2.3.1 Experiment 1

First, we looked at how well our content-based models predicted the link structure of the title graph. The rationale of this experiment depends on the observation that the existence of a link between two titles is a strong hint of being related because at least one user of İTÜ SÖZLÜK conceived an association between them. We treated this problem as a binary classification task. Title pairs with a link in between were labeled as positive while the pairs which do not have a link between themselves were labeled as negative. The models classified the instances and predicted their labels by applying a threshold on the estimated relatedness scores. In order to avoid the burden of deciding on thresholds and to compare the characteristic behavior of the models in more detail, we constructed the receiver operating characteristic curves (ROC) and compared the area under the curve (AUC) measures of the models.

2.3.2 Experiment 2

The fact that İTÜ SÖZLÜK has an evolving graph structure means that the set of links in a given snapshot does not determine the entire set of related titles. In terms of the binary classification problem, the dataset we constructed contains many false negatives –title pairs which do not have a link in between but otherwise related to each other and maybe will be linked to each other in the future. Thus, the performance of a model cannot be measured solely by using the graph structure because even if a model perfectly discriminates the linked pairs from the unlinked ones this does not mean that it can correctly predict the future links. These concerns about the validity of AUC measures motivated a second experiment based on human-collected data.

[4]http://en.wikipedia.org/

In this experiment, we randomly sampled 100 titles from the set of target titles and call these sample targets *items*. Then for each item, we obtained three –or whatever number the model can recommend at most– recommendations per model. For instance, if each of the 5 models provided 3 recommendations for a given item then we got 15 item-recommendation pairs to be scored. Then the pairs were randomly divided into 11 roughly equally sized subtests –we employed a factorial design to balance the distribution of item-method pairings through subtests. The participants in the experiment were 25 voluntary İTÜ SÖZLÜK users. All of the 25 participants rated the item-recommendation pairs on the 11^{th} subtest on a scale from 1 to 4, defined as, 1 totally unrelated, 2 somehow related, 3 related and 4 very much related. Then each of the remaining 10 subtests was randomly assigned to one of the participants and was rated in a similar way. In short, each participant rated two subtests and each subtest was rated by at least 2 participants. If the participants did not know the meaning of any of the titles in a pair, then they were asked to either give a blank score for the pair or to look them up in İTÜ SÖZLÜK before deciding on the score.

3. RESULTS AND DISCUSSION

3.1 Experiment 1

The performance measure obtained from the first experiment was AUC of the models in the binary classification task. The obtained ROC curves are given in Fig 1. In this experiment, we evaluated the term-based model (TRM) and the concept-based model (CON). The back-link model (BCK) was left out in this analysis because it already uses the link information between the titles. The referral-based model (REF) was also left out because the relatedness scores calculated by REF was very sparse and it was not possible to construct a meaningful ROC curve for it (the AUC calculated for REF was 64%, slightly above random performance).

TRM clearly outperforms CON and obtains an almost perfect AUC of 94.60%, while CON obtains an AUC of 84.66%, which is a still high on its own.

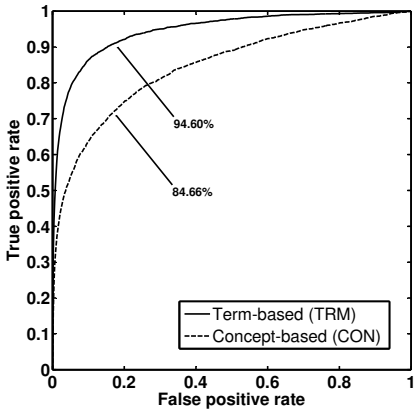

Figure 1: ROC curves of term-based and concept-based models on link prediction task. The AUC values are indicated for the corresponding curves.

As it can be seen in Fig 1, both TRM and CON are quite

successful in terms of predicting the existing links. However, our initial aim was to improve upon the link structure because the graph is very sparse and a good recommender system must also find related titles even if there are no links between them. To evaluate this aspect, we carried out the second experiment.

3.2 Experiment 2

For the 100 stems sampled from the set of all target titles, a total of 1217 recommendations were made. TRM, CON, and RND provided 3 recommendations for all stems; REF provided 2.84 recommendations per stem on the average and BCK was able to provide only 0.33 recommendations per stem on the average.

On the 11^{th} subtest, which is rated by all 25 participants, the average pairwise Spearman rank correlation among the participants' ratings was 0.56 suggesting a significant albeit moderate inter-rater reliability.

If we group the recommendations according to the models we can calculate the average score per item for each of the 5 models. The results are given in Fig 2 along with the 95% confidence intervals on the means. There are no significant differences between BCK, TRM, and CON ($p > 0.1$ for a 2-sample t-test). However, they all perform significantly better than REF and RND ($p < 0.001$).

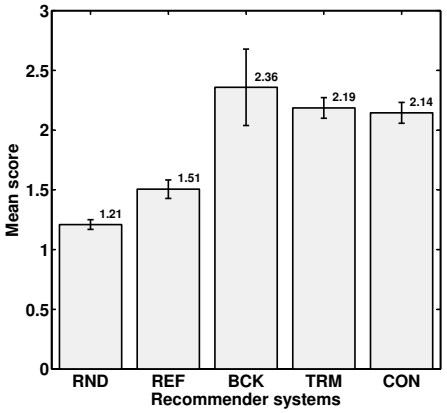

Figure 2: Average human ratings per item-recommendation pair. The error bars indicate 95% confidence intervals and average ratings are given at the top of the bars.

The performance of BCK has a higher variance than of TRM and CON. This is probably due to the fact that it produced fewer recommendations. This observation justifies our initial motivation: BCK might be producing high-quality recommendations, but since the link structure is very sparse, we cannot obtain even one recommendation for a majority of the titles.

4. CONCLUSION

Increasing the coverage is an important problem for the recommendation systems designed for domains with heavy-tailed distributions. We showed that in our case, both the titles placed on the heavy tail of the degree distribution (i.e. old but forgotten titles) and the newly created titles (i.e. cold start problem) suffer from the low coverage of graph-based models. However by using content-based models, we obtain at least three times more recommendations per title without any significant decrease in recommendation quality. Since the choice of three recommendations per item in our analysis was an arbitrary decision, the observed difference between graph-based and content-based models in terms of coverage is at best underestimated.

5. ACKNOWLEDGMENTS

Authors would like to thank Çağatay Gürtürk, the owner of İTÜ SÖZLÜK, for his support during the data fetching; Marco Baroni and Şule Gündüz Öğüdücü for their valuable feedback on a draft of this manuscript.

6. REFERENCES

[1] D. Adams. *The Hitchhiker's Guide to the Galaxy*. Pan Books, London, 1979.

[2] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[3] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[4] M. Göker and C. Thompson. The adaptive place advisor: A conversational recommendation system. In *Proceedings of the 8th German Workshop on Case Based Reasoning, Lammerbuckel, Germany*, pages 187–198, 2000.

[5] A. Gunawardana and C. Meek. Tied boltzmann machines for cold start recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 19–26. ACM New York, NY, USA, 2008.

[6] A. Herdağdelen, E. Aygün, and H. Bingol. A formal treatment of generalized preferential attachment and its empirical validation. *EPL (Europhysics Letters)*, 78(6):60007+, 2007.

[7] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[8] S. Pado and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

[9] Y. Park and A. Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 11–18. ACM New York, NY, USA, 2008.

[10] M. Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm University, Faculty of Humanities, Department of Linguistics, 2006.

[11] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523, 1988.