Recommender system based on user-generated content

© Turdakov Denis

Institute for System Programming, RAS
turdakov@gmail.com
Ph.D. adviser: Kuznetsov S.D.

Abstract

Recommender systems apply statistical and knowledge discovery techniques to the problem of making recommendations during live user interaction. This paper describes a novel approach of building recommender systems for the Web with the aid of user-generated content. Recently certain communities of Internet users have engaged in creating high quality peer reviewed content for the Web. In our approach we are planning to extract the semantics of such user-generated content and to use these semantics to make more useful recommendations.

1 Introduction

1.1 Recommender Systems

Recommender systems attempt to predict items (web pages, movies, books) that a user may be interested in, given some information about the user's profile.

Collaborative filtering is the most popular approach to building such systems. Individual users are automatically joined in groups based on similarity in their interests or past behaviour and recommendations are made based on the preferences of their group members. Another method in generating recommendations is based on predicting users' interests based on his/her past preferences. In the former approach new content is compared against user's past preferences and similar items are recommended. Both systems suffer from a number of deficiencies: mainly they both fail to recommend novel interesting topics. For example, a recommender system for travellers based on collaborative filtering can assign a user to a class of people who visit European capitals. However, once he visits all of the capitals, this system cannot recommend anything new to this person. More generally, it has been described in [1] that both types of recommender systems that strive to achieve maximum accuracy in classification do not lead to useful recommendations.

In our work, we avoid this difficulty by generating recommendations of Web pages with the aid of semantics, extracted from user-generated content. This allows us to make recommendations based on the relationships between concepts, created and peer-reviewed by a large community of users. It is obvious that building a recommender system for Web pages that extracts semantics from all of the content on the Web would be very resource demanding. Instead we picked Wikipedia as our source of user-generated semantics, which is a comprehensive and up-to-date corpus of knowledge and relationships between concepts.

1.2 Wikipedia

User-generated content refers to various kinds of content that is produced or primarily influenced by end-users. However, average quality of Web content is quite poor, as evidenced by vast amounts of Web spam and unverified information submitted by non-authoritative users. Therefore, instead of using the Web, we chose Wikipedia as our base, since it is a body of user-generated content that is all encompassing and at the same time of high quality and peer reviewed.

In every article of Wikipedia links guide users to associated articles, often with additional information, and lists of categories for each article organize Wikipedia articles in a taxonomic structure. These links convey important semantic information that we can use to produce high quality recommendations. Furthermore, any Internet user is welcome to add further information, cross-references, or citations, so long as user do so within Wikipedia's editing policies and to an appropriate standard. So after a time semantically rich and high quality peer reviewed content emerges.

The English-language Wikipedia currently (when article was written) contains more then 1,500, 000 articles (6,000,000 when including redirects, discussion pages and portals).

Furthermore, we can consider that Wikipedia is a form of web summarization because of its broad scope, conciseness and ability to quickly reflect new trends. Therefore, Wikipedia can provide us with extremely useful information about articles and Web page relationships.

2 Problem formulation

The main goal of this research is to develop a recommender system for the Web based on user-

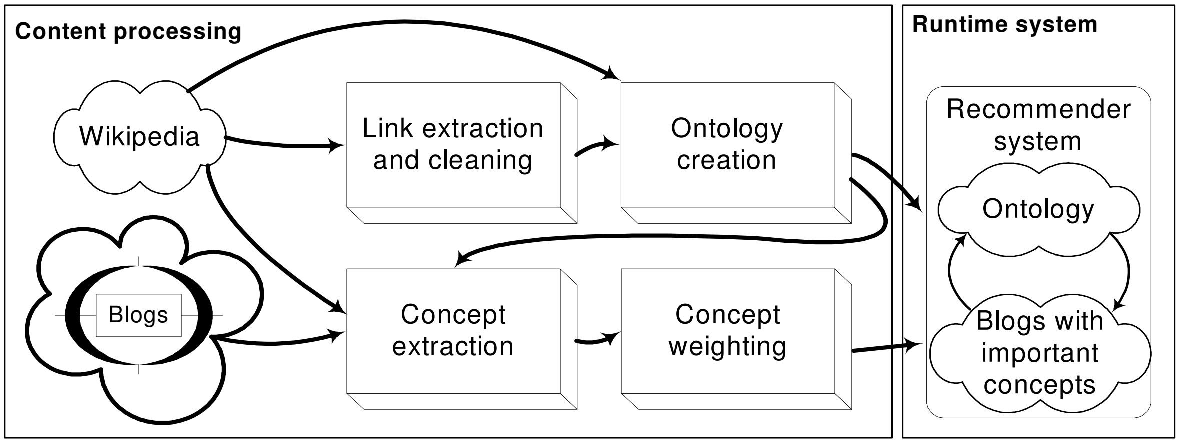

Figure 1: Recommender system architecture

generated content and novel semantic techniques. Instead of recommending web pages or sites, our system will recommend blogs. This choice stems from a number of considerations. First, blogs are the most dynamic part of Internet and are constantly getting renewed. Secondly, blogs have simple structure in comparison to web sites, and it's easier to evaluate a recommender system for blogs versus complex web sites. Finally, crawling a representative subset of the web is a daunting task.

Web log recommender system would have substantial practical value, since this type of content search is a difficult task for the user. For example, discussion about new products starts long before official announcement of these products. But since the user doesn't know about this product, he cannot perform a meaningful search.

The overall system architecture is presented in figure 1. Within the content processing framework there two key processes: Wikipedia analysis (top of figure) and blogs processing (bottom of figure). We analyze Wikipedia and create an ontology based on its structure. Then we extract and rank concepts from the blogs, making use of the ontology. Also in this stage we associate blogs and ontology through concepts. For example, our system associates a blog with keywords "Moscow" and "Capital of Russia" with an article about Moscow in Wikipedia. Then we use these associations to find blogs similar to users preference set (set of blogs that characterizes users interests) with the aid of the ontology.

In thee next two sections we focus on two main stages of the system: the first one is Wikipedia link cleaning and ontology extraction; second is establishing semantic relationships between Wikipedia concepts and web logs. In the following sections we describe the remaining stages.

2.1 Cleaning Wikipedia links

Wikipedia has its own markup, links to internal and external articles, redirects and list of categories. All of this information would be useful for our research. So far we are only using article titles and internal links. Though this is only a small part of information could be extracted from Wikipedia, we would be able to get results very quickly and rate their quality.

When you analyze the link topology of Wikipedia carefully, you will notice that in many cases an article will contain links to other articles that are completely unrelated. Thus, for example, in article about Moscow there is a link to an article about Fahrenheit temperature scale. Clearly, we should make a distinction between these kind of links and high quality links such as link from "Moscow" to "Capital of Russia". So we need a mechanism to clean or rank links on the basis of their quality.

For solving link cleaning task it is necessary to investigate how such low-quality links appear. Typically, Wikipedia editors carefully insert relevant links between key concepts of their article to other articles. Occasionally a rogue user will insert a bunch of irrelevant links into an otherwise quality article. We can see a similar pattern with Web spam, where spammers create large artificial chunks of the Web to boost the page rank of some specific site. Therefore we would like to modify and use emerging Web spam combating algorithms [2] to clean Wikipedia links. In order for these methods to be applicable we need to make sure that Wikipedia has the same properties.

Widely known models of the evolution of the Web [3, 4] describe global properties such as degree distribution or the appearance of communities. These models indicate that overall hyperlink structure arises by copying links to pages depending in their existing popularity. For example in the most powerful model [4] pages within similar topics copy their links that result in "rich gets richer" and we see power law degree distribution where the exponent vary approximately from 2 to 3.

So, web graph relates to the class of scale-free networks with most distinguishing characteristic are that their degree distribution follows a power law relationship. The second property of this class of networks is self-similarity: a large-enough supporter set should behave similar to the entire Web. Thus we can guess that properties of Wikipedia links graph and its subgraphs would be same to the Web graph.

The basic idea is to analyze rank distribution of some page in its neighborhood. If link distribution in

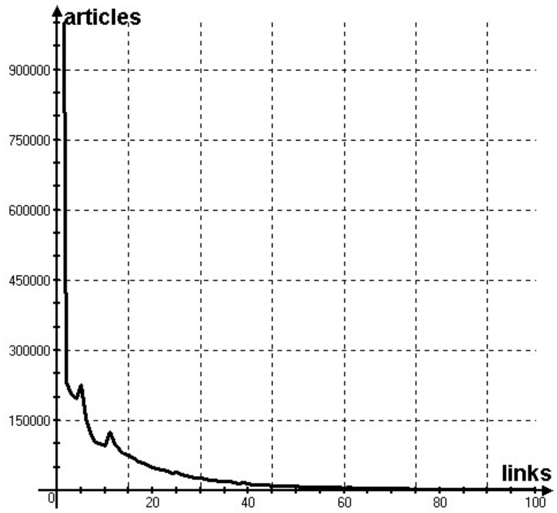

Figure 2: Wikipedia link distribution

some article neighborhood isn't power law we have a high probability that page rank is artificially overstating. Therefore emerging spam detection algorithms require the following properties:

- Power-law link distribution
- Self-similarity

We have analyzed Wikipedia and found that its link structure follows the power-law distribution (figure 2) and it follows that self-similarity holds for Wikipedia. Hence we can use modified Web spam detection algorithms for this task.

2.2 Ontology extraction

A naïve way to produce recommendations is to recommend blogs associated with nearest neighbors in the Wikipedia link graph. However there are serious problems with this method. Researchers [5] proved that uncorrelated power-law graph having the exponent approximately from 2 to 3 will also have ultrasmall diameter d ~ ln ln N (for Wikipedia d = 2.75). For our work it means we can't use only the link structure to make recommendations, since we will end up recommending the whole collection of blogs. So we need to extract additional knowledge from Wikipedia that will help us select a few relevant links for recommendations. Therefore the second stage of Wikipedia processing deals with extracting semantic information from link and articles. Next, we give a short overview of ontologies used in information retrieval and describe our ontology model.

Semantic extraction and ontology development is well-studied topic. WordNet is the most successful hand crafted semantic lexicon for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining glosses (Definitions and/or example sentences).

Simple methods are used in IBM research center to make automatic semantic annotation of WEB pages. *Seeker* is a platform for large-scale text analytics, described in [6]. *SemTag*, an application written on the platform to perform automated semantic tagging of large corpora. Authors use small and simple TAP ontology [7] to process large amounts of web pages. This is the largest scale semantic tagging effort to date. We work with much smaller data; therefore we can use a more complicated method and extract more semantic data from a document.

In recent works researches have started to extract semantics from Wikipedia; the most profound work was done by Kozlova [8]. She extracts an ontology with structure similar to WordNet. To evaluate the quality of the ontology she compared the performance of the ontology-driven classification of Reuters collection with extracted ontology versus WordNet, and achieved better results.

In her work article link structure and article structure itself was used for ontology extraction. For example, if analyze the link [[Capital of France | Paris]] we can easily produce synonyms: "Paris" and "Capital of France. Also, if a document is linked under one of the special sections like "see also", "similar topics" it indicates, that this document has something to do with the topic.

Unlike the previous works we will extract a more semantically rich ontology. For now we will use categories, "see also" links and general links in the articles.

List of categories form a directed graph over the articles of Wikipedia, which can be very useful in pruning irrelevant links when making recommendations. For example, Wikipedia article about Kurchatov contains links to "physics", "Physico-Technical Institute" and other topics that are poor recommendation candidates. With the aid of categories we can prune these links and recommend more relevant topics such as articles about Kurchatov colleagues. This is the most basic use of our ontology; we will investigate more sophisticated methods in our future work.

2.3 Web logs processing

Now we deal with preprocessing web logs. In order to find blogs most similar to user preference set it's necessary to extract terms from each blog and correlate these terms with concepts from the ontology. This will enable us to make recommendations based on these concepts.

When we correlate blogs with concepts each blog will become associated with a large number of concepts. In order to identify essential concepts we use a modified tf-idf weighting scheme [9]. We avoid recomputing idf every time the blog collection is updated by computing idf using only Wikipedia.

2.4 Generation of recommendations

At runtime the recommender system derives top-N recommendation from the ontology based on users preference set. Little research has been done on this topic. An ontology-based information retrieval model [9] exploits ontology-based knowledge bases to improve search over large documents. This approach includes an ontology-based scheme for the semi-automatic annotation and retrieval of documents. We plan to use and extend this technique for computing similarity between blogs and ranking recommendations.

3 Related work

Tapestry [10] is one of the earliest implementations of collaborative filtering based recommender systems. This system relied on the explicit opinions of people from a close-knit community, such as office workgroup. However, recommender system for large communities can't depend on each knowing others. Later on several rating-based automated recommender systems were developed. The GroupLens research system [11] provides a pseudonymous collaborative filtering solution for Usenet news and movies. Ringo and Video Recommender are email and web-based systems that generate recommendations on music and movies respectively. A special issue of Communications of the ACM [12] presents a number of different recommender systems. Although these systems have been successful in the past, their widespread use has exposed some of their limitations such as the problems of sparsity in the data set, problems associated with high dimensionality and so on.

A myriad of other recommender systems exist, particularly on e-commerce sites. Schafer [13] examines and categorizes a large set of these commercialized recommender systems. In addition, numerous recommenders in a variety of domains have been developed for research purposes, including MovieLens (films), Ringo (music), and Jester (jokes).

All of these systems are based on collaborative filtering. Correspondingly they have problems as stated above. We avoid these problems by using high quality user-generated content as a foundation for making recommendations.

4 Conclusion

In this article we propose a novel architecture for building recommendation systems and formulate the major directions of future work. In the future we plan to modify and apply Web spam detection algorithms for cleaning Wikipedia links. We will then evaluate various approaches to making recommendations using the extracted ontology (we have given a basic example of such an approach in Sections 2.2 and 2.4).

Finally, we will implement a complete blog recommender system based on the described techniques and evaluate it on the Internet users.

6 References

[1] S.M. McNee, J. Riedl, and J.A. Konstan. "Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems". In the Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006), Montreal, Canada, April 2006.

[2] Andras a. Benczur, Karoly Csalogany, Tamas Sarlos Mate, and Uher. SpamRank – Fully Automatic Link Spam Detection, work in progress, Computer and Automation Research Institute, Hungsrian Academy of Sciences, 2005.

[3] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the word-wide web. Physica A, 281:69–77, 2000.

[4] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In Proceedings of the 41st IEEE Annual Symposium on Foundations of Computer Science (FOCS), 2000.

[5] R. Cohen and S. Havlin, Scale-free networks are ultrasmall, Phys. Rev. Lett. 90, 058701 2003.

[6] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. IBM Almaden Research Center, 2003.

[7] TAP ontology page. http://ontap.stanford.edu/

[8] Natalia Kolova. Automatic ontology extraction for document classification. Computer science department, Saarland University, 2005.

[9] David Vallet, Miriam Fernández, and Pablo Castells. An Ontology-Based Information Retrieval Model. Universidad Autonoma de Madrid.

[10] Goldberg, D., Nichols, D., Oki, B. M., and Terry,D. Using Collaborative Filtering to Weave an Information Tapestry. Communication of ACM, 1992.

[11] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. GroupLens: An open Architecture for Collaborative Filtering of Netnews. Proceedings of CSCW, 1994.

[12] Resnik, P., and Varian, H. R. Recommender Systems. Special issue of Communication of the ACM, 40(3), 1997.

[13] J. Schafer, J. Konstan, and J. Riedl. Electronic commerce recommender applications. Data Mining and Knowledge Discovery, Jan. 2001.