Coopeer: A Peer-to-Peer Web Search Engine Towards Collaboration, Humanization and Personalization

Jin Zhou, Kai Li and Li Tang
Department of Automation
Tsinghua University, Beijing, China
{zhoujin00,li-k02,tangli03}@mails.tsinghua.edu.cn

Abstract

Most centralized web search engines currently become harder to catch up with the growing step of people's information need. Here, we present a fully distributed, collaborative peer-to-peer web search engine named Coopeer. The goal of the work is to complement centralized search engines to provide more humanized and personalized results by utilizing users' collaboration. Towards this goal, three main ideas are introduced: (a)PeerRank to use cooperation among users for evaluation, (b)query-based representation to obtain more humanized description about documents, and (c)semantic routing algorithm to obtain user-customized results.

1. Introduction

It has become increasingly difficult to search for useful information on the web, due to its large size and unstructured nature. Researchers have developed many different techniques to address this challenging problem of locating relevant web information efficiently. The most conventional example is Centralized Search Engine (CSE).

One major problem with CSEs is that they do not facilitate human user collaboration, which has potential for greatly improving web search quality and efficiency. Another major problem with CSEs is that they ignores completely the interests and preferences of users. Different users will be answered with a same list of results for a same query. However, moving from a centralized paradigm towards a distributed one, brings in some advantages which cannot be exploited earlier. Briefly, they are related to the fact that information has been collected, selected, stored and shared among users according to their interests.

In this paper, we propose a peer-to-peer (P2P) approach for web searching, implemented in a system named Coopeer. In Coopeer, information about web pages and user searching experiences is shared in a peer-to-peer fashion. Our proposal attempts to create a highly distributed system where each node stores a part of the web model used for indexing and retrieving web resources in response to queries. All users share these partial models that globally create a consistent model for the web resource that is equivalent to its centralized counterpart.

2. System Overview

At the beginning, we give a description to the work flow of Coopeer. Launching a search run, the requestor forwards queries based on local indices which record the semantic content of remote peers. One may adjust the local index in terms of her own preference. Receiving a remote query, the local peer issues a local information retrieval. To facilitate the work, the query-based representation is introduced to index documents. Based on query representation, cosine similarities between documents and next queries can be given. The documents whose similarities exceed a certain threshold are considered to be relevant enough to queries. Receiving the returned results, the requestor ranks documents by considering both preference of local user and feedback from remote companions with PeerRank algorithm.

2.1. PeerRank

Enlightened by collaborative filtering and social voting, we present a novel PeerRank algorithm working in P2P network. In fact, when a user raises a new request, it is likely that the same job has been ever done by a lot of other users for many times. The searching experience, such as ever used query terms and the evaluation of results, would be much helpful to the new requestor. However, neither term-frequency method[5] nor linkage method[2] utilizes the human searching experience. In part this is due to that

the highly centralized search engines prefer those machine-based methods. By contrast, in the Coopeer network, all the users are taken as a "Referrer Network". PeerRank determines page's relevance by examining a radiating network of "referrers". Documents with more referrers gain higher ranks. In short, PeerRank has following advantages:

(a) Collaboration of users. PeerRank has potential to obtain better rank order, as collaborative evaluation of human users is much more precise than description of term frequency or link amount.

(b) Prevention of spam. PeerRank makes a great improvement to prevent spam, since it is difficult to pretend evaluation from human users.

2.2. Query-based Representation

Coopeer uses a novel type of query-based representation based on the relevant words introduced by human users with a high proficiency in their expertise domains. This is due to two thoughts that (a)the human users' queries are more accurate to describe the retrieved documents and (b)people tend to use the same subset of words very frequently. In this thought, each peer maintains a piece of query-based inverted index. According to query-based inverted index, cosine similarity between new query and documents can be computed. In contrast with content-based representation[1],query-based fashion is more humanized. Meanwhile, query-based invert index avoid most work in content-based inverted index, such as term breaking, stopwords removing and etc. This is benefit to local P2P software and reducing user workload.

2.3. Semantic Routing Algorithm

Coopeer uses a semantic routing algorithm based on directed flooding manner. Each peer maintains a Topic Neighbor Index to describe the content of neighbor peers. Topics in the index are only interested in by local peer. A peer only forwards a query to those promising peers whose content is more relevant to the query. There are twofold advantages in semantic routing algorithm:

(a) Semantic Expression of Content. Coopeer client can forward queries based on the semantic content of neighbor peers.

(b) Self-organized User Community. The local index of a peer is updated by those passing responses if their topics are similar enough to one of local topics.

Other P2P systems[3][4][6][7], consider file similarity in terms of a key space generated by a cryptographic hash. Users must know a file's key in order to retrieve it from the network. By contrast, Coopeer is there for those situations in which users don't know exactly which file they want.

3. Conclusions

This paper outlines a P2P search engine called Coopeer towards collaboration, humanization and personalization. There are three main features in Coopeer: (a)The integration of the users' subjective ratings to rank the results in a collaborative fashion; (b)The introduction of query-based representation for web pages by query terms instead of the elements in web contents; (c)Insertion of personalized factor for searching results by routing in self-organized user community. These advantages are only possible in P2P network where the information and cost is shared among all the members.

4. Acknowledgements

We would like to thank the subjects at the University of Tsinghua who participated in user study. The first author would also like to thank Xiao-Long Zhu and Yun-Gang Zhang for many insightful comments on this draft.

References

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[3] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 161–172. ACM Press, 2001.

[4] A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, pages 329–350, 2001.

[5] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[6] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan. Chord: a scalable peer-to-peer lookup protocol for internet applications. *IEEE/ACM Trans. Netw.*, 11(1):17–32, 2003.

[7] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141, UC Berkeley, Apr. 2001.