LaBeeB: Systematic Peer Clustering for Building a Semantic Peer-to-Peer Web Search Engine

Motaseem Al Amour

Abstract. Search engines (SE) were only considered in presenting the results of a search query ranked according to a certain algorithm. This means all users will see the same results for the same query (at the same time). Although these users differ in many aspects, this fact was not considered in modern search engines. Peer-to-Peer SEs have tried to imitate centralized ones, with little success. They were faced with massive amount of data, very dynamic structure of the environment and a big number of peers. The real power of a p2p network was not utilized sufficiently, which is the peers themselves. Peers are a representation of human identities in the internet. A peer (or a human) can be categorized by the following criteria: the language it speaks, the country it comes from, the age group it belongs to and the human character it falls under. This very peer in turn has many interests, and so it will visit web pages that match these interests. This information for this peer and million others will be then saved in the p2p network. With proper interpretation of this information a semantic web search engine can be built and a systematic method can be used to rank the result of a query according to the number of visited web pages visited by peers that have the same criteria as the query initiator.

In this paper we present LaBeeB\(^1\), an innovative p2p web search engine that can resolve user queries effectively in a semantic fashion and can then rank the results based on human factors.

1 Introduction

To enable LaBeeB to rate web pages to the number of visits of peers similar to the query initiator, one should be able to measure the similarity between peers or

\(^1\) An Arabic word means knowledgeable and intelligent.
humans in general. We define similarity as a combination of language, country, age and personality or character which are known for each peer and are called peer’s attributes. We assume peers who share these attributes are very alike. And if these peers also share the same interest in politics for example, then we argue when one of these peers visits a page with a political topic, most probably other similar peers will be interested in that page too. The answer is to be found in [19] and [20] where the term limbic character is defined. It focuses on our unconscious brain (the limbic system) and how it is affecting our decisions, behavior and all day activities. That part of the brain is responsible for 75% of our decisions [19]. Three basic emotions balance, dominance and stimulation are defined; the combination of these emotions gives one of six limbic characters for every person.

Most approaches for a p2p content search have suggested a Distributed Hash Table (DHT) that holds a global inverted index of (keyword, web address) with some improvements like [?] [23] [24] [25] [21]. This design will deliver poor performance for a p2p environment with millions of users, billions of web pages and terabytes of data. Others proposed a semantic approach [8] [7], by adopting a hybrid p2p network where some tasks are assigned to special peers. LaBeeB will not use any dedicated peers because first it is not a wise choice to put the load of a p2p SE on some super peers that they could easily be overloaded and second we will use a true p2p network where the load is distributed across all peers.

Although a DHT provides an effective and guaranteed data retrieval, it has the disadvantage of a large maintenance cost to keep it consistent, especially in a highly fluctuating dynamic environment. That is why [18] argues that the cost for maintaining the DHT outweighs the advantage of having one in the first place as most time is spent on updating DHT indices. LaBeeB will overcome this disadvantage by first reducing the amount of data to be stored in the DHT by adopting another form of the global (keyword, web address) inverted index, and second by clustering peers and so having not one but several DHTs. With this design a DHT size is reduced drastically and thus the maintenance cost is minimized.

Text Categorization or Text Classification (TC) [14] [15] [16] is the task of automatically sort a set of documents (in our case a web page) into one or more categories based on the document’s content. Thus TC is a method to categorize texts semantically. TC is formalized by the function \( \phi : D \times C \rightarrow True, False \) where \( D \) is a set of documents and \( C \) is a predefined set of categories. LaBeeB categorizes web pages using TC techniques, by using this principle a web page is indexed by its category rather than by all keywords it contains. This means we can replace a global huge (keyword, web address) inverted index with a smaller semantic (category, web address) one.

The rest of the paper is organized as follows. Sections 2 presents the p2p search engine architecture. Section 3 illustrates peer clustering. We conclude with a discussion of what we have accomplished.