Self-Adaptive Ontology based Focused Crawler for Social Bookmarking Sites

Aamir Khan, GLA University, Mathura, India
Dilip Kumar Sharma, GLA University Mathura India

ABSTRACT

It is not possible for one person to explore or surf all the relevant websites pre-training to his/her topic. A user might not be able to get the results that he/she expects from the search engine but another user might have some knowledge about some website containing the information about the first user’s topical query. Users share their information on a common sharing platform known as SBS (Social Bookmarking Sites). In SBS a user posts a question seeking some knowledge about a certain topic, and then the people who have some knowledge about any website related to the query topic post the URLs of the website. This paper presents a novel method to verify the authenticity and validity of the URL posted in the SBS. The performance of our method is further increased by using a dictionary based learning methodology that finds the contextually similar words that are added to the Ontology.

KEYWORDS

Crawler Policies, Focused Crawling, Ontology, Self-Adaptive, Social Bookmarking Sites

INTRODUCTION

The internet has become an important part of our day-today life. Total internet users in 2014 were approximately 2.91 billion, and this number is increasing day by day. As the internet users have increased so have their requirements, and to cope up with this the size of World Wide Web (WWW) also increased accordingly. The number of websites in 1995 was a mere 65 million but in 2014 it exponentially rose up to 970 million in number. Despite having so many websites user browses only a small percentage of the websites are relevant to him/her due to a number of reasons. The reasons may include:

1. Low recall of the search engine algorithm/system,
2. Websites having some synonymous/similar names,
3. Websites in different languages, and
4. Non-indexed websites (either new or old). Since many of the websites are not indexed by any of the search engines and are not indexed in the deep web, they show no or barely link to the indexed websites or web pages that is why it is difficult to access them even if they contain information that is more relevant to the user than the indexed pages D. K. Sharma and A. K. Sharma (2011).

That is where Social Bookmarking Sites (SBS) come to our aid. SBS are centralized in nature and it allows users to store and share internet bookmarks. SBS also allow the user to annotate, add, edit, and share the web document’s bookmarks. Suppose a user posts some need of a particular website...
selling artefacts from mainland China, then other active users give the information about the first user’s need or provides the link to the website(s) selling the product.

Most traditional crawlers impose a heavy communication loads due to the fact that they use the improper ontology for query processing as stated in D. K. Sharma and A. K. Sharma (2009). Focused crawlers are different from traditional crawlers as they satisfy some specific predicates towards the crawl frontier and thereby analysing and maintaining the hyperlink exploration process. For example, a crawler’s main objective might be only to extract pages from the ‘.ac.in’ domain. The probability of the page being relevant to the user query topic is being calculated by the focused crawler before downloading the webpage, whereas a traditional or standard crawler downloads the webpage irrespective of their topic relevance. Focused crawlers extract the thematic words pertaining to the content with the help of either dictionary based or other methods.

The very first time the focused crawler structure was proposed by Chakrabarti et al. (1999). Focused crawler’s general architecture is as shown in Figure 1 instead of searching the complete web, only a specific domain is searched by the focused crawler. There may be many factors on which the search process of a focused crawler depends upon, but it is broadly classified into two categories. First, the area of interest of the user and Second, the predefined set of topics already present with the search engine. A focused crawler which is specifically a topic driven, i.e. it retrieves web documents or web pages belonging to a particular topic group are known as topic driven crawler.

The focused crawling process is dependent upon two counterparts: Classifier and Distiller Chakrabarti et al. (1999). A classifier is used to calculate the significance of the search topic and the document retrieved, i.e. this module of focused crawler is used to classify between the relevant document and non-relevant document. A distiller is used to search for the valuable access points by using fewer numbers of links that guide to a massive number of suitable documents, i.e. the relevant access point (path to reach form source to destination in least hops or in least time) from the complete web graph is explored.

From the list of seed URLs, the state of the URL is checked, i.e. whether a URL is crawled or not. If a URL is not previously crawled, then the crawler selects the URL and loads the page and the outgoing links from the webpage are extracted. Then the extracted links are then stored in the

Figure 1. Generalized Framework of a focused crawler
Geo-Tagging News Stories Using Contextual Modelling
[www.igi-global.com/article/geo-tagging-news-stories-using-contextual-modelling/186825?camid=4v1a](www.igi-global.com/article/geo-tagging-news-stories-using-contextual-modelling/186825?camid=4v1a)

Role-Based Multi-Agent Systems
[www.igi-global.com/chapter/role-based-multi-agent-systems/28077?camid=4v1a](www.igi-global.com/chapter/role-based-multi-agent-systems/28077?camid=4v1a)