Journal of Computer Science 6 (4): 450-456, 2010
ISSN 1549-3636

An Adaptive Updating Topic Specific Web Search System Using T-Graph

Ahmed Patel
Department of Computer Science, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, The National University of Malaysia,
43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract: Problem statement: The main goal of a Web crawler is to collect documents that are relevant to a given topic in which the search engine specializes. These topic specific search systems typically take the whole document's content in predicting the importance of an unvisited link. But current research had proven that the document's content pointed to by an unvisited link is mainly dependent on the anchor text, which is more accurate than predicting it on the contents of the whole page. **Approach:** Between these two extremes, it was proposed that Treasure Graph, called T-Graph is a more effective way to guide the Web crawler to fetch topic specific documents predicted by identifying the topic boundary around the unvisited link and comparing that text with all the nodes of the T-Graph to obtain the matching node(s) and calculating the distance in the form of documents to be downloaded to reach the target documents. **Results:** Web search systems based on this strategy allowed crawlers and robots to update their experiences more rapidly and intelligently that can also offer speed of access and presentation advantages. **Conclusion/Recommendations:** The consequences of visiting a link to update a robot's experiences based on the principles and usage of T-Graph can be deployed as intelligent-knowledge Web crawlers as shown by the proposed novel Web search system architecture.

Key words: Topic specific search engines, DDC, T-graph, web crawling, web robot

INTRODUCTION

Modern Web-based search systems are classified as centralized with a single point of control and distributed with either centralized control or distributed autonomous management control. These systems operated by Web robots can be broadly classified into two categories:

- Non-topic specific Web robots
- Topic specific Web robots

Non-topic specific Web robots are initially supplied with seed URLs acting as starting points for fetching the documents by the Web crawler which contain many links to other documents but typically have no or very little useful content information without the example documents at hand. Example documents can be as simple as keyword lists that specify the crawler's topic focused or non-focused scope as they have to index as many Web documents as possible without regard to their topic. Unfortunately such Web coverage leads to many problems, some of which are listed below:

- They are not easily scalable. Even if they are scalable, all the scaled components are dependent on a single underlying centralized database (O'Meara and Patel, 2001).
- There is no definitive metric or ranking to identify the importance of the fetched documents and to store all of them found in the crawling paths. Hence, any simple keyword search query results in rendering thousands of documents ranging from high, little or no relevance to be delivered by the search. The quality of "high" relevance is dependent on the keyword search query count rather than any sophisticated algorithm or metric.
- The WWW is continuously growing at millions of pages per day and nearly 600 GB of text content changes every month (Chakrabarti *et al*., 1999). Thus, it becomes a daunting task to index the whole WWW at the same time maintaining the indexed documents freshness or currency.
- The whole search system needs to be under one proprietary control only and can provide search services to only publicly indexable Web.

The above disadvantages can be eliminated by dividing the whole WWW into specific topics and retrieving the Web resources/documents relevant to specific topics. Such division of WWW gives rise to topic specific search engines crawling their way and

harvesting one or more specialized topic(s). They are specialized in one or more topic(s) and collect documents which are relevant to their specialized topic(s) and discard the irrelevant documents as they go about doing their business. Each topic specific search engine can be owned by independent owners and can interact with other topic specific Web search systems. Mutual interest and viable benefits can be achieved through agreeing the demarcation of specialization topic boundaries. As the indexing tasks are distributed among independently owned search systems, the burden of currency maintenance is also distributed and is more easily performed unlike in non-topic specific search systems. Similarly, the private Web can also be searched and indexed through a system of autonomous, federated and cooperative distributed topic specific search systems (O'Meara and Patel, 2001). The idea behind this is to generate revenue by providing search engines which specialize in topics and subjects like Gaelic football, soccer and netball.

Some advanced topic specific Web robots use the concept of machine learning techniques to identify the relevant documents hidden by surpassing the irrelevant documents (Baldi *et al*., 2003). Typically the topic is represented by example documents which contain the keywords belonging to the specialized topic.

While the topic specific search systems can address the deficiencies of the non-topic specific search systems, they are still in their infancy. Most of the present topic specific system architecture depends on the analysis of the whole document content in predicting the importance of an unvisited link rather than the text surrounding the unvisited link. Passerini *et al.* (2001) showed that predicting importance of unvisited link based on the anchor text is more accurate than prediction based on the whole page/parent pages content. Hence we present in this study an architecture, which will analyze the text present only within the topical boundary of an unvisited link in determining its importance. Here the topical boundary will be the words, which surround an unvisited link, including the words of the link itself and differ greatly from other words/phrases in the document with respect to their topics. To find the topic boundaries around an unvisited link, Dewey Decimal Classification (DDC) is exploited.

The remainder of this paper discusses the concepts of DDC and topic boundary detection methods, T-Graph's technique and its use, analysis and evaluation. This is followed by a presentation and discussion of the proposed novel Web search system architecture and the paper is concluded by giving an indication of extensions and future research work.

MATERIALS AND METHODS

Dewey decimal classification: The Dewey Decimal Classification (DDC) system (Frank, 2010; OCLC, 2010) is the world's most widely used library classification system used as a general knowledge organization tool that is continuously revised to keep pace with latest knowledge, either semantic or otherwise. At the broadest level, the DDC is divided into ten main classes, as listed below which together cover the entire world of knowledge:

- Generalities
- Philosophy and psychology
- Religion
- Social science
- Language
- Natural science and mathematics
- Technology (applied sciences)
- Arts
- Literature
- Geography and history

Each of the above classes has ten divisions. These divisions are further divided and sub-divided. In this way, the Dewey classification system progresses from the general to the specific. Digits following the decimal point after the first three digits are used to make the classification even more specific. The number shown beside the topic represents the unique real number associated with a particular topic. For example, the deduction of topic "butterfly" is shown in Fig. 1.

From now onwards we refer to the real number associated with any topic as the D-Number (read as Dewey Number). Any topic should have exactly 3 digits before the decimal point. If the most significant digit before the decimal is zero or the least significant digit after the decimal point is zero then the zeros should not be neglected as done in ordinary decimal numbers.

Exploiting DDC: DDC is a hierarchical classification method. The unique feature of DDC is D-Number. This D-Number can be utilized to determine whether two words/phrases belong to the same/related topic or not.

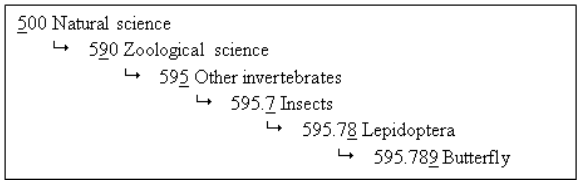

Fig. 1: Catching the topic "butterfly"

For example let us suppose that three words 1, 2 and 3 have D-Numbers 505.4443, 505.4445 and 303.88 respectively. By observing the D-Numbers we can easily determine that words 1 and 2 belong to the same super topic, but word 3's super topic is entirely different. Computationally this can be achieved by comparing the D-Numbers digit-by-digit starting from the most significant digit. If the most significant digit of two D-Numbers is same, then both the words belong to the same topic at the broadest level.

In the act of determining the topic boundary around an unvisited link the following modifications for the DDC are performed:

- Each D-Number associated with topic is represented as a string.
- The D-Numbers are modified in such a way that the trailing zeros are eliminated if the decimal point is not present in the D-Number. For example the list of broad topics and their D-Numbers shown in Fig. 1 are modified as 0 for Generalities, 1 for Philosophy and Psychology and so on. So, a topic's D-Number need not contain 3 digits before the decimal point, provided the decimal point is not present.
- All the words in the DDC are stemmed by using Porter Stemming algorithm (Porter, 2009).

The first modification allows us to represent the most significant zeros and least significant zeros after the decimal point intact. The second modification will make the D-Number representation ideal for the outlier detection. The third modification will enable us to compare the document words with DDC words perfectly. For example, if in some moment the system needs to know the D-Number associated with word Butterflies and the DDC System contains word Butterfly rather than Butterflies then any string comparison function will report a mismatch despite the singular form of the word "butterfly" is present in the DDC. Such mismatch errors can be eliminated by exploiting the concept that all forms of a word (singular, plural, verb, adverb, noun and adjective) will have the same root word. The Porter Stemming algorithm can be applied to the words in DDC System and document words so that correct word comparison is made in determining the D-Number associated with a word.

RESULTS

Topic boundary: Diligenti *et al*. (2000) have proposed a topic specific search system architecture, which is based on the concept called context graphs. The context graphs store the contexts so that any retrieved document can be mapped to nodes of context graph and the system can predict the number of documents that are to be retrieved to get the related/target document. However, the whole document content is taken into account while determining the matching node. The novel Web search system proposed in this paper was inspired by the concept of Diligenti *et al*. (2000). It mainly depends on the comparison of words within the topic boundary around an unvisited link with the nodes of Treasure Graph (T-Graph). Thus, it finds the matching node(s) and determines the number of documents that are to be downloaded to get the most relevant documents. The Watchdog component will observe the consequences in visiting a link to update the nodes content of T-Graph, if necessary. The T-Graph concept not only determines the number of documents that are to retrieved, but it also guides the system which links in the retrieved documents that are to be followed to get the most topic related documents.

The whole concept right from the construction of T-Graph to the complete operation of the system is dependent on the detection of topic boundary around an unvisited link. Starting from the words of an unvisited link, a 2-D graph is plotted with the D-Number and its length as the dimensions. A special document parser is used which segregates the document into words/phrases. It gives phrases if any phrase is directly present in DDC, else it gives the words. Each word/phrase can have more than one D-Number. For example, the word clothing belongs to several disciplines. The psychological influence of clothing belongs in 155.95 as part of the discipline of psychology; customs associated with clothing belong in 391 as part of the discipline of customs; and clothing in the sense of fashion design belongs in 746.92 as part of the discipline of the arts. As a result for each word/phrase, more than one point may be plotted. The plotting of the points is stopped as and when the paragraph boundary is reached. If the unvisited link is a list item, then also the plotting is stopped after plotting all the list items. Then the plotted points are analyzed. Such pair of points at which plotting is stopped and analysis is conducted, are called break points (Initially the break points could constitute the starting and ending of the paragraph itself). At some moment of analysis the dots seem to deviate much from the dots of the words corresponding to the link. If all the dots corresponding to words between break points are thickly populated around the words related to link, the plotting is repeated by considering the new break point pairs. They constitute the beginning of the preceding

paragraph and ending of the succeeding paragraph that add words before and after paragraphs of the present paragraph. Figure 2 shows an example of how a graph looks when plotted along with the explanation of the concept as follows:

- All the dots constitute the words between break points.
- Each light colored (unbolded black or grayish) dot shows the point corresponding to the word, which is not the anchor text.
- Each dark colored (bolded black) dot shows the point corresponding to the text comprising the unvisited link.
- For each word/phrase, more than one point may be present. But words which are thickly populated in the graph are considered more likely to be a hit. This is shown as (A) encircled in red in the graph in Fig. 2. This thickly populated group of words in the graph is typically called a galaxy.
- As the points become sparse the plotting is stopped and the words present in the galaxy are considered as the group of words that are present within the topic boundary around an unvisited link.
- Note that all the dots correspond to the words between two break points, called a break point pair. As it can be observed that only words present in the galaxy are considered and not all the words between the break point pairs. Break points constitute the paragraph boundaries. They just specify the words which are to be plotted before analyzing the graph.

Hence the topic boundary is nothing but group of words. The above process simply explains how to determine the topic boundary around an unvisited link manually. Computationally we can get such galaxy of words by first performing outlier analysis (Han and Kamber, 2006). Outliers are words in the perspective of our topic concept) that do not comply with the general behavior or model of the data, that is, they deviate from the words within the topical boundary around an unvisited link from the perspective of the topic.

Outlier analysis is used by detecting the outliers between break points and eliminating the outliers from the words within the same break points, which constitute the galaxy. There are many ways in Data Mining subject field which obtain the outliers from a sample data set, using such methods and algorithms as Statistical outlier detection, Distance based outlier detection and Deviation based outlier detection (Kargupta *et al*., 2008), which are beyond the scope of this study.

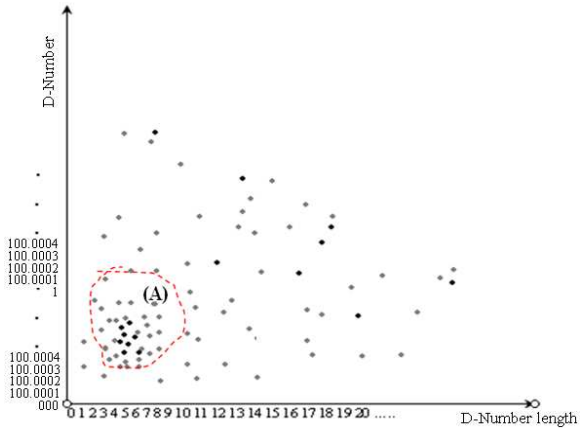

Fig. 2: An example graph of D-Numbers and their length

T-Graph: The concept of T-Graph is similar to the concept of Context Graph (Diligenti *et al*., 2000) except that each node contains the data that typically occurs within the topic boundary around a hyperlink. Each node in the T-Graph contains data under the following five attributes:

- The Immediate Sub-section Heading (called ISH) component.
- The headings of the subsections/sections which contain the ISH component.
- The Main Heading (MH) component.
- The Data Component (DC).
- The Destination Information Component (DIC).

The ISH component's data is compared with the sub-heading's (say U) words containing the unvisited link, to obtain a similarity measure say sim_{ISH}. The sub-headings (if any) containing the U are compared with SH data to obtain a similarity measure of sim_{SH}. The main heading of the document containing the unvisited link is compared with MH component to obtain a similarity measure of sim_{MH}. The words found within the topical boundary of the unvisited link are compared with the words present in DC to obtain a similarity measure of sim_{DC}. The DIC component contains the information such as the least number of documents that are to be downloaded to get the target documents. Thus, the Overall Similarity Measure (OSM, which is a function of sim_{ISH}, sim_{SH}, sim_{DC}, sim_{MH}) for each node and the words present within the topical boundaries of an unvisited link is calculated. The nodes whose OSM is above a threshold value are considered. Once the matching nodes are selected, their respective DIC' s data are analyzed to get the least number of documents

that are to be downloaded and depending on that value the priority of the link is calculated. If none of the node's OSM value is above the threshold, then the link is least prioritized, but still visited. In this case a "watchdog" component or system facility observes the consequences of visiting such a link and updates the robot's experience in the T-Graph structure table. The OSM of the unvisited link is a function of sim_{ISH}, sim_{SH}, sim_{DC}, $sim_{MH,}$ defined as:

$$OSM = f(sim_{ISH}, sim_{SH}, sim_{DC}, sim_{MH})$$

The sim_x is the cosine similarity (Baldi *et al*., 2003) and can be calculated as:

$$sim_x(v_x, v_y) = \frac{v_x \bullet v_y}{|v_x||v_y|}$$

Where:

v_x = The term frequency vector, in which the subscript x assumes ISH, SH, MH and DC.

v_y = The term frequency vector of the words, in which y assumes the subheading words containing the unvisited link for comparison with ISH, the subheadings containing subheading which contains the link for comparison with SH, the main heading of the document for comparison with MH component's data and the words within topic boundary of an unvisited link for comparison with DC.

This primarily constitutes a very efficient and effective function for the algorithm.

Figure 3 shows an organization of a T-Graph. One of the nodes is elaborated to show the components in the node. All the nodes are divided into various levels. The top-most level of the node depicts highest number of documents that are to be downloaded. However, it can also happen that a node present at a higher level can directly point to the lowest level, such as nodes 1 and 3 shown in Fig. 3. In such cases the document pointed by the unvisited link is given high priority as it may contain the link which will directly point to the lowest level nodes. Consequently, the T-Graph acts as a road map for downloading the relevant documents as rapidly as possible.

For the organization and presenting the nodes in the T-Graph, the following rules apply:

- A node is assigned to a level x, if and only if it contains a pointer/link to at least one node at level x-1, regardless if the node may also contain any number of links to other nodes of the lower levels (x-2, x-3…).
- Only one link exists between two nodes at different levels.
- No two nodes can be combined into one. No link can exist between nodes at the same level.

From the above, it can be observed that the whole concept of T-Graph is based on the logical assumption that any unvisited link's importance is solely dependent on the text present within the topical boundary around it. It also assumes that if a link is present as text in the DC component of any node, it will typically point to the pages containing the links, which are again represented as nodes. There are some other nodes of which the system detects and if followed will result in non useful documents. Such nodes are called dead nodes. For example, see node 4 in Fig. 3 that illustrates this fact. If the text within the topical boundary of a link matches with such a node, the link is discarded by the system and moves to the next matching node.

T-Graph construction: Initially some sample data is collected which constitutes the documents with links after they have been filtered against dead nodes.

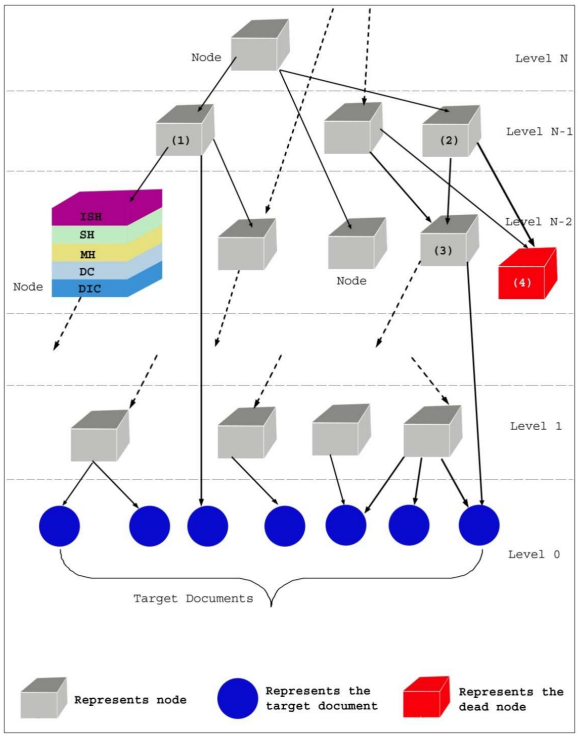

Fig. 3: T-Graph structure

This data set acts as the training data. The topical boundary around each link present in e ach document is found and a node is constructed and the document pointed by the link is retrieved and analyzed.

This procedure is repeated until a desired number of nodes at desired number of levels are formed. The working of the graph is tested on known documents and it's predictability (for example, which can be defined as a percentage of correct predictions or some other criteria) is measurable. If the desired prediction is not achieved, then the graph construction is repeated. Once the construction phase is completed, the graph is ready for usage. However, the graph construction or modification is a continuous on-going process. The watchdog component or system facility (shown in Fig. 4 as Watchdog process within the T-Graph subsystem) closely monitors the new situations or occurrences and updates the T-graph. This system becomes adaptive and involves both supervised learning (starting with initial graph building) and unsupervised learning (maintenance or enhancement of the graph) based on the new situations as and when they are encountered.

Proposed novel Web search system architecture design using T-Graph: Figure 4 shows the overall architecture of the new system using T-Graph. The Crawler component locates the documents pointed by the links present in the Fetcher Queue. For each document fetched it places the response in the response queue. The response queue contains the documents or HTTP response. If the page cannot be downloaded due to temporary disconnection from the Web and/or non-freshness of the link, the system maintains the current HTTP response details to perform the calculations based on these criteria. The results are used, regardless whether the page is modified or not and without downloading because any re-connection or freshness is regarded as a minor hazard. This effectively ensures continuity of the process.

The Response queue processor processes the documents fetched and analyzes whether the document belongs to the specialized topic or not. It also utilizes the T-Graph, considered as the system's experience and calculates the importance of the unvisited links to the document. Depending on the number of documents required to download, it assigns the priority to the link and places in the Fetcher Queue. If none of the nodes present in the T-Graph matches with the words present within the topical boundary of the unvisited links then the link is assigned some lower priority and placed in the Fetcher Queue accordingly. The fetched document pointed by such link is carefully analyzed by the

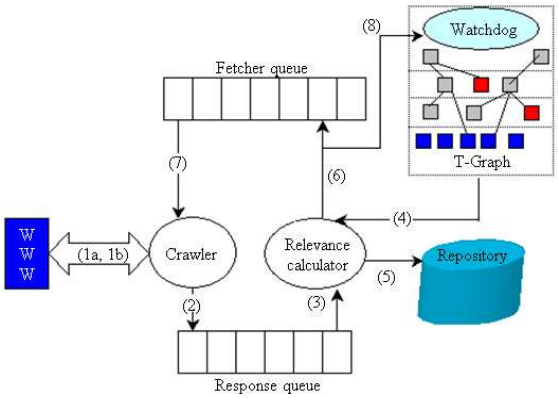

Fig. 4: Novel web search system architecture design

Watchdog process and it cyclically updates the T-Graph to incorporate new knowledge and experiences gained within it repertoire of operations. The priority of the links present in the queue is incremented periodically in order to avoid starvation, which is called aging.

Novel web search system architecture design with the following sequence of events (Fig. 4):

(1a) Crawler sending the request for document fetching from the Web corresponding to the link present in the fetcher queue
(1b) Crawler acquiring the document
(2) Crawler placing the fetched document in the Response Queue
(3) Relevance calculator selecting the fetched document from the response queue
(4) Relevance calculator getting the information from T-Graph for calculating the priority of the link present in the document;
(5) Relevance calculator placing the document information in the repository, if it finds that the document belongs to the specialized topic
(6) Relevance calculator placing the link extracted into the fetcher queue
(7) Crawler selecting the link to be fetched from the fetched queue
(8) Watchdog analyzing the response for updating of the graph

DISCUSSION

The evaluation and use of the T-Graph principle: The notion of using T-Graph and building a new Web search system is based completely on the concept of topic boundary detection around an unvisited link. During the running of the system, the evaluation of this principle is performed first. For this purpose, the system is supplied with various documents containing different discussions related to different topics and the results are

compared with the manual detection of topic boundaries in the same document set. After this is evaluated, the concept and use of T-Graph is tested and evaluated. For this purpose, a set of known documents coming in the path for downloading a target document are taken and analyzed to assess whether the system is properly following the correct/optimal path or not.

The overall performance of the system from the perspective of precision is then measured by taking another topic specific search system, which considers the whole document content into account and which works on entirely different concept. Both the systems are supplied with same keyword list and seed URLs and the number of documents downloaded is reached to an observable predefined precision.

It is beyond the scope of this study to give all the runs, analysis and full evaluation of the novel T-Graph based Web search system. This will be done in a follow-up paper in the future.

CONCLUSION

This study has presented and discussed the concepts and principles of T-Graph and the novel Web search system architecture which is based on the analysis of text which is present in the topical boundaries surrounding an unvisited link, unlike taking the whole document's text into consideration. It is argued that the text within the topical boundaries around an unvisited link will only decide the content of the document pointed by the unvisited link. This shortcoming, which is the focus of this study, shows how a refinement of the concept of T-Graph guides the Web crawler to follow the optimal path to reach the target document as rapidly as possible. It shows that if the text contained in the Data Component (DC) of any node contains a link then it will typically point to the document containing the links which match its child nodes.

The concept of T-Graph also enables the system to work on the concept of association analysis for mining the words that most frequently co/re-occur. The function for calculating the OSM is not over-emphasized because the relative priority of the text within a topic boundary, the immediate sub-heading, other sub-headings, heading of the document cannot be same for all the documents. If one were to take the OSM as verbatim at its face value, the context of finding the most appropriate document for the topic with a given set of keywords could be misleading and the ultimate value of the system will be defeated. To overcome this problem, an investigation is being conducted to assess a more advanced adaptive approach in which the system intelligently acquires the relative importance that can be modified as the document is being analyzed and the T-Graph updated accordingly. This is being performed together while investigating the Machine Learning and Language Modeling approaches for efficient topical boundary detection around an unvisited link. To this end, the prototype system presented in this study is designed in such a way that any new efficient method for boundary detection can be plugged and played into the system without disturbing the overall system's architecture. This is useful for testing new techniques and algorithms and performing various experiments.

Finally, these topic areas of T-Graph and Web-based search systems are exciting from theoretical, experimental and practical points of view. The T-Graph concept has many useful practical applications.

REFERENCES

Baldi, P., P. Frasconi and P. Smyth, 2003. Modeling the Internet and the Web: Probabilistic Methods and Algorithms. John Wiley and Sons, Chichester, England, Hoboken, New Jersey, ISBN: 0470849061, pp: 285.

Chakrabarti, S., M. van der Berg and B. Dom, 1999. Focused crawling: A new approach to topic-specific Web resource discovery. Proceedings of the eighth international conference on World Wide Web, Toronto, Canada, pp: 1623-1640.

Diligenti, M., F. Coetzee, S. Lawrence, C.L. Giles and M. Gori, 2000. Focused crawling using context graphs. Proceedings of 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt, pp: 527-534.

Frank, 2010. Let's Do Dewey. http://frank.mtsu.edu/~vvesper/dewey2.htm

Han, J. and M. Kamber, 2006. Data Mining-Concepts and Techniques. 2nd Edn., Morgan Kaufmann, ISBN: 1558609016, pp: 800.

Kargupta, H., J. Han, P. Yu, R. Motwani and V. Kumar, 2008. Next Generation of Data Mining. Taylor and Francis/Chapman and Hall/CRC, USA., ISBN: 9781420085860, pp: 601.

OCLC., 2010. Dewey Services at a glance. Online Computer Library Center, OCLC, http://www.oclc.org/dewey/

O'Meara, T. and A. Patel, 2001. A topic-specific Web robot model based on restless bandits. IEEE Internet Comput., 5: 27-35.

Passerini, A., P. Frasconi and G. Soda, 2001. Evaluation methods for focused crawling. Lecturer Notes Comput. Sci., 2175: 33-39.

Porter, M., 2009. An algorithm for suffix stripping. http://tartarus.org/~martin/PorterStemmer/index.html