TELKOMNIKA, Vol.10, No.1, March 2012, pp. 189~198
ISSN: 1693-6930
accredited by DGHE (DIKTI), Decree No: 51/Dikti/Kep/2010

CT-FC: more Comprehensive Traversal Focused Crawler

Siti Maimunah[1], Husni S Sastramihardja[2], Dwi H Widyantoro[2], Kuspriyanto[2]
[1]Faculty of Information Technology, Surabaya Adhitama Institute of Technology,
Jl. A.R. Hakim 100, Surabaya, Indonesia, +62315945043/+62315994620
[2]School of Electrical Engineering and Informatics, Bandung Institute of Technology,
Jl. Ganeca 10 Bandung, Indonesia
e-mail: s.maimunah@yahoo.com, husni@informatika.org, dwi@if.itb.ac.id, kuspriyanto@yahoo.com

Abstrak

Saat ini semakin banyak orang yang menggantungkan kebutuhan informasinya pada WWW, termasuk para profesional yang harus menganalisis data sesuai domainnya untuk memelihara dan mengembangkan usahanya. Sebuah analisis data tentunya membutuhkan informasi yang komprehensif dan relevan terhadap domainnya. Untuk kebutuhan aplikasi ini digunakan focused crawler sebagai agen pengindeks informasi Web yang relevan terhadap topik tertentu. Dalam usaha meningkatkan presisinya, ternyata focused crawing menghadapi permasalahan rendahnya nilai recall. Hasil studi terhadap karakteristik struktur hyperlink dalam WWW menunjukkan bahwa banyak dokumen Web yang tidak terhubung secara langsung melainkan melalui kositasi dan koreferensi. Focused crawler konvensional yang menggunakan strategi forward crawling tidak memungkinkan untuk mengunjungi dokumen dengan karakteristik tersebut. Penelitian ini menawarkan sebuah kerangka penelusuran yang lebih komprehensif. Sebagai pembuktian, CT-FC (focused crawler dengan kerangka penelusuran baru) dijalankan pada data DMOZ yang representatif terhadap karakteristik WWW. Hasil eksperimen menunjukkan bahwa strategi ini mampu meningkatkan recall secara signifikan.

Kata kunci: *focused crawler, kositasi, koreferensi, recall*

Abstract

In today's world, people depend more on the WWW information, including professionals who have to analyze the data according their domain to maintain and improve their business. A data analysis would require information that is comprehensive and relevant to their domain. Focused crawler as a topical based Web indexer agent is used to meet this application's information need. In order to increase the precision, focused crawler face the problem of low recall. The study on WWW hyperlink structure characteristics indicates that many Web documents are not strong connected but through co-citation & co-reference. Conventional focused crawler that uses forward crawling strategy could not visit the documents in these characteristics. This study proposes a more comprehensive traversal framework. As a proof, CT-FC (a focused crawler with the new traversal framework) ran on DMOZ data that is representative to WWW characteristics. The results show that this strategy can increase the recall significantly.

Keywords*: focused crawler, co-citation, co-reference, recall*

1. Introduction

The rapid growth of information makes general search engine more difficult to provide services effectively. On general search engine, users must select and open the document first before determine whether the information list is relevant to their needs. This job can be time-consuming and tedious for the user [1]. Instead of general search engine, several professional organizations need domain search engine to meet their information needs. This domain search engine indexes only documents relevant to specific topics. To index the information domain search engine uses focused crawler as an agent to traverses WWW and downloads documents relevant to the specified topics.

Focused crawler must determine which link to visit to maximize relevant documents obtained and avoid links that are not important to minimize irrelevant documents. A good strategy is needed to determine the seed pages in an effective and predictions on which ones deserve a link is followed to obtain relevant documents before the actual download [2]. For now

Received May 11th, 2011; Revised October 27th, 2011; Accepted January 9th, 2012

conventional focused crawler can only reaches relevant documents that are connected by downloaded documents out-links. Actually there many characteristics of relevant documents hyperlink structure in WWW and some relevant documents could not be obtained by other relevant document out-links. Thus, need a new strategy to ovoid locality search trap in focused crawling [2].

2. Related Works

Web crawler is a program that utilizes the Web graph structure to move from one document to others in order to obtain Web documents and add them or their representation to a local storage media. Thus, the crawling process can be viewed as a graph search problem.

In its simplest form, search process of crawling system starts from a seed URLs and then by using downloaded document out-links to visit other URLs. This process is repeated by increasing out-links that are generated from new documents. The process will end if the number of documents considered is sufficient or meets certain criteria. In general, the infrastructure of the crawling process shown in Figure 1(a) [3].While building a general search engine, many problems will be encounter such as the need of huge resources particularly in terms of providing storage and bandwidth for crawling process and services to various domain users.

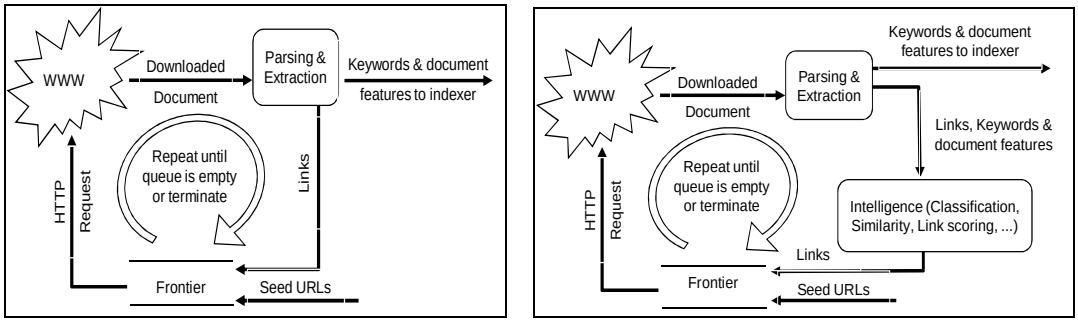


Figure 1. Many relevant documents that are connected through (a) co-citation documents and (b) co-referenced documents

To index information related to specified topics, focused crawler has a smart component to determine the search strategy on the web graph. This component leads to the graphs that are relevant to a particular topic. Figure 1(b) shows focused crawling infrastructure. Qin [4] categorizes the proposed focused crawling methods into two parts: Web analysis algorithm and Web search algorithm. Web analysis algorithm used to assess the relevance and quality of Web documents and Web search algorithms used to determine the optimal order in which the target URLs are visited.

First generation crawler based on traditional graph algorithms, such as breadth-first search or depth-first search [5]. From set of seed URLs, the algorithm follows hyperlinks leading to other documents recursively. The main objective of crawling system is to search over the Web and download all documents found. Thus, the material contained in the document content will be least observed.

Instead of general crawler, a focused crawler must obtain Web documents relevant to a particular topic efficiently. Generally, researchers proposed Web content-based search strategy. This strategy is derivation of text retrieval that already has a mature theoretical base. Salton [6] proposed a vector space model that represents each document or query by a vector. In this model, each term represents a single dimension and the weight that accompany to each dimension represents the term contribution related to document material. Furthermore, the lexical representation can infer the semantic meaning of a document by using lexical topology. Based on the model Rijsbergen [7] provided a hypothesis, i.e.: a document with the same vector space to a relevant document will have a high probability of relevance. Search engines have used the lexical metric traditionally to rank any documents according to their similarity to query [8].

One direction of Web document hyperlinks (out-links) make focused crawler search limits to top-down strategy, called forward crawling. Actually many Web documents are organized in tree structure. When the focused crawler is in a leaf position, this makes serious obstacle to find highly structured or sibling/spouse relevant documents. For example, when focused crawler find a computer science researcher main page from a hyperlink of paper list at a conference site, it needs a good strategy for crawling other members' documents of computer science department. Without hyperlink to the other department members' documents explicitly, conventional focused crawler will not be able to move up to the department main page and to the other members' documents. This condition makes conventional focused crawling recall low.

3. Conventional Focused Crawling Precision and Recall Trade-Off

There is a trade-off between precision and recall of conventional focused crawling. Higher conventional focused crawler result precision, make the recall getting lower. Figure 2(a) shows focused crawling process that ignores irrelevant documents. Thus, the crawling result has low precision but high recall. On the other hands Figure 2(b) shows focused crawling process that avoid irrelevant documents can increase precision and declining the recall. This is because of WWW characteristics, which permits many relevant documents, connected to the others indirectly.

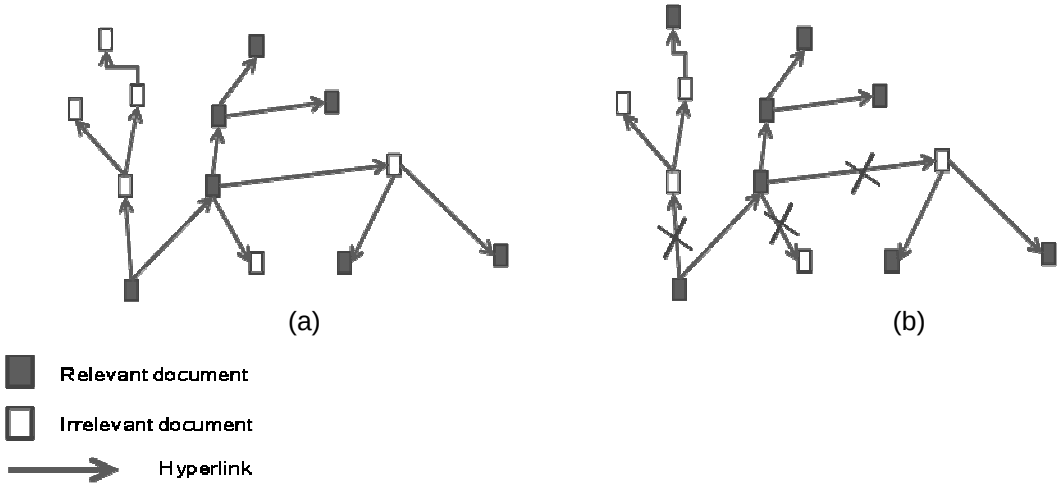


Figure 2(a) When the goal of crawling system is just higher recall, it will download all of documents both relevant and irrelevant ones until all relevant documents are downloaded; (b) Focused crawling system cuts the link through irrelevant documents to maintain precision

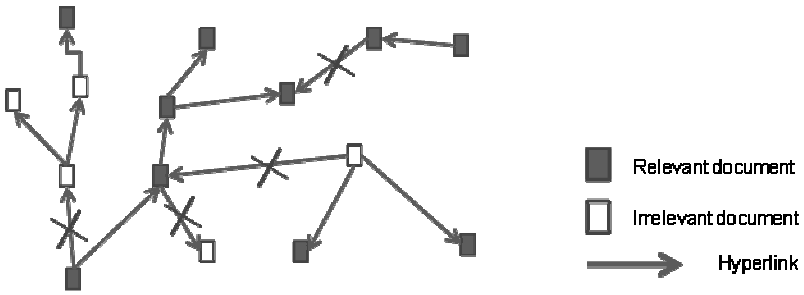


Figure 3. Conventional focused crawler may not reach relevant documents connected through co-citations or in-links of downloaded documents

Many relevant documents also connected to the others through co-citation documents or by in-link of downloaded documents that make conventional focused crawler has low recall (Figure 3). The following chapter describes more detail about WWW characteristics.

4. WWW Structure Characteristics

In general, Web graph characteristics identified by previous researchers categorized into four quadrants of Cartesians diagram (Figure 4). Horizontal axis describes connecting type between relevant documents (directly/indirectly) and vertical axis describes search direction that must be done to obtain the relevant documents (forward/backward).

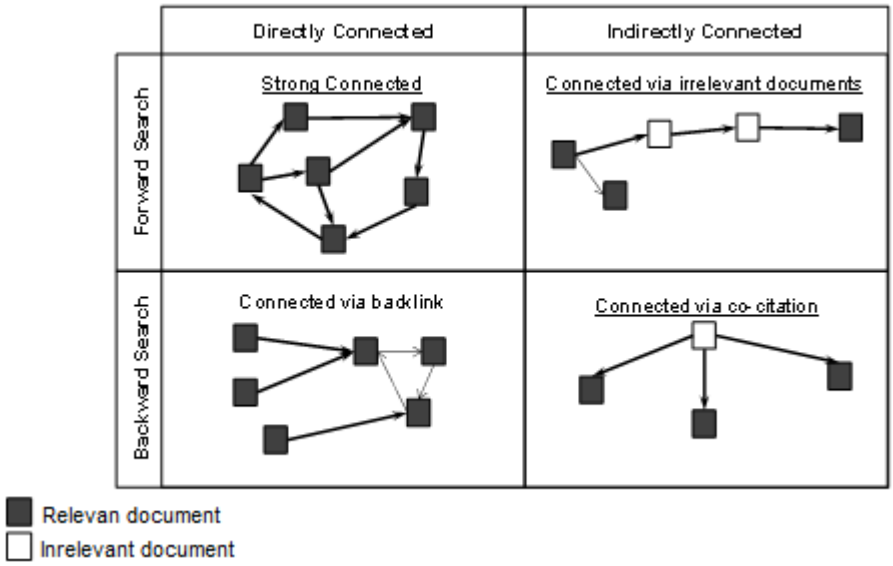


Figure 4. Four WWW characteristics quadrants

Relevant documents in quadrant I which are connected directly and in a forward direction search, have strong connected characteristic, i.e. there are connections from one document to others and there is a cycle in the inter-links graph. Quadrant II (connected indirectly and in a forward direction search) contains relevant documents, which have indirectly connected characteristic, i.e. connected through one or several irrelevant documents [9], [10], [11]. Relevant documents in quadrant III connected directly through in-links of downloaded documents. Quadrant IV contains relevant documents, which are connected via co-citation documents [12], [13].

5. Focused Crawling System

Focused crawlers considered as a Web information searcher agent. User query initiates the information search. The user query expressed in the form of seed URLs relevant to specified topic. Afterwards, focused crawler downloads documents related to the seed URLs and maps them to an appropriate concept. The concept mapping is to understand queries concept given by user and limits the Web retrieval fields.

An ontology can be useful to know the relationships between concepts. Combination of query concept and the available general ontology used to set up local ontology of specified topic. If a concept has no link to the query concept then the concept should be removed from the local ontology. Finally, each lexicon related to each concept in the local ontology can be used as a reference to assess the documents' relevance. Figure 5 shows the focused crawling framework.

Measuring parameters and measurement criteria may be used to assess system optimality. These parameters and the measurement criteria may influence to system design. Therefore, the following sub-chapter will discuss the measuring parameters and measurement criteria before discussing focused crawling system in details.

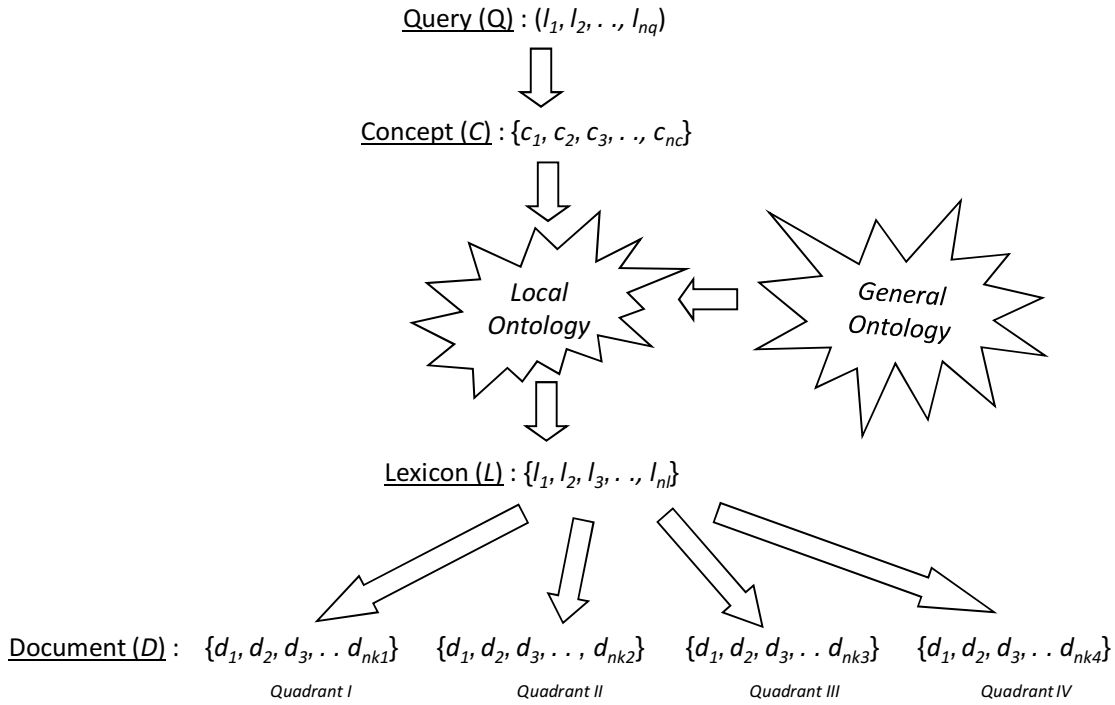


Figure 5. Focused crawling framework

5.1. Increasing Focused Crawling Precision

Focused crawler's operation is initialized by a given seed URLs. The operation runs according to downloaded document's relevance assessment and the obtained links. The relevance assessment is based on query as an abstraction of seed URLs.

To increase precision, focused crawler uses semantic analysis to asses document relevance. The semantic analysis is to obtain the desired topic concept (Figure 6(a)). As an illustration, when a user wants a pet topic. Let documents related to the seed URLs talk about cat babies. Keyword 'cat baby' is acquired at the pre-process. In syntactic analysis, the result documents may contain words 'baby' and/or 'cat'. Based on this syntactic analysis results, documents that contain the word 'baby' will be set true even though the document discusses about a human baby. Meanwhile, focused crawler will reject a document containing word 'dog' because it does not contain the word 'baby' or 'cat'.

There are two disadvantages in syntactic search: (1) By taking the documents containing the word 'baby' without considering what kind of baby will make more irrelevant documents are downloaded. This condition will reduce precision. (2) When crawler rejects any documents which are not contain the words 'baby' or 'cat' even though the document is in the same concept, it will make the recall becomes low.

If a query has one major concept (hereinafter referred to as topic), then focused crawling result has high precision because the query does not have multi meanings (polysemy). The greater number of concepts related to the query, implies that precision decreases exponentially (prediction accuracy is 1/|c|) (Figure 6(b)). Thus, to increase the precision, the query must be mapped onto exactly one major concept or topic (Q:C = 1:1).

5.2. Increasing Focused Crawling Recall

To increase focused crawling recall, local ontology of query concept has to generate after determining the query concepts. Local ontology is generated by the main query concepts substitution into available general ontology and trim the related concepts of the same topic. Figure 7 illustrates the substitution process. Lexicon list and its combination derived from the local ontology to assess document's relevance. The derivative results also include synonyms of the main topics lexicon. The completeness of the topic's concepts and synonyms knowledge may influences the increase of the focused crawling recall.

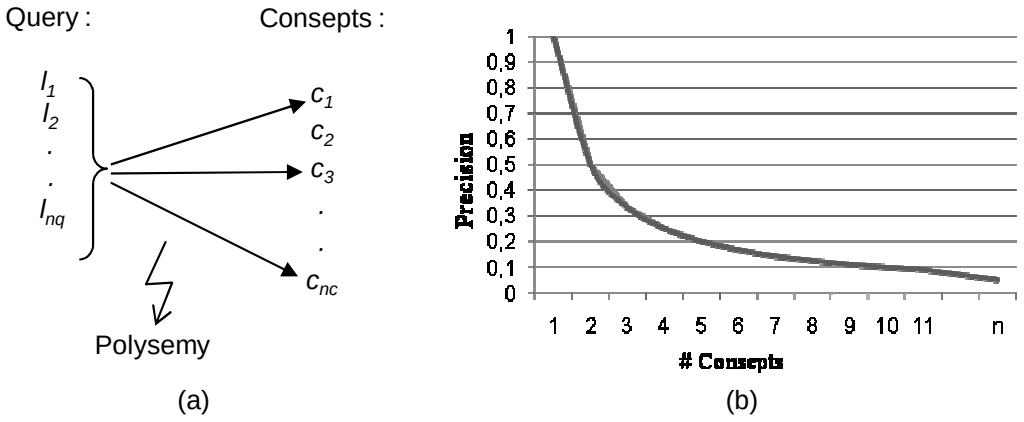


Figure 6. (a) Query to concept mapping. If the mapping produces more than one concept, then there is a polysemy or ambiguity in query meaning; (b) Precision decreases in accordance with the number of query concepts

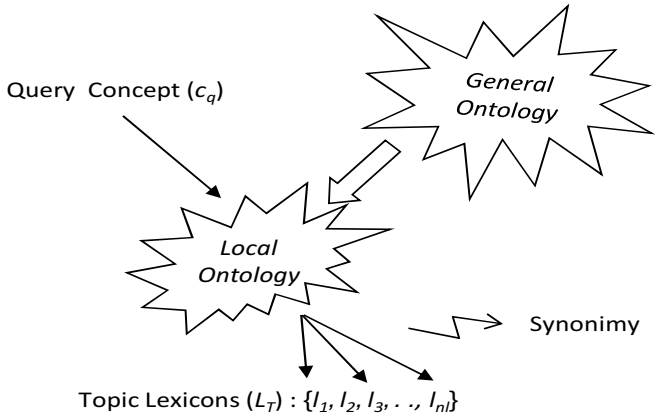


Figure 7. Main query concepts substitution into general ontology to generate local ontology and associated lexicon list of topic

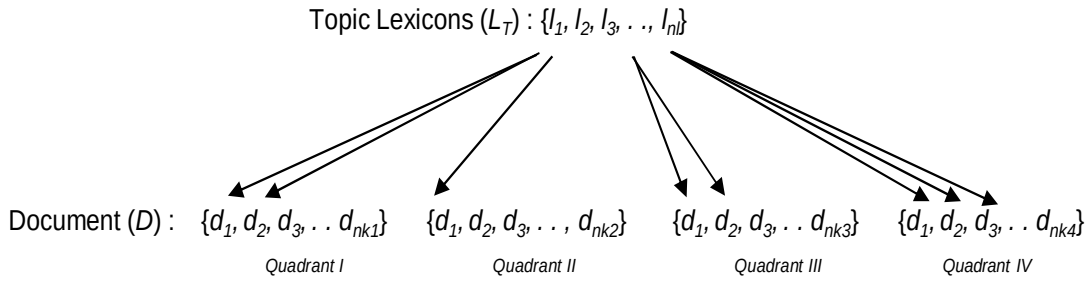


Figure 8. Four quadrants of focused crawling search spaces.

Based on the WWW characteristics, beside the two variables above (related concepts and lexicon synonyms), completeness of exploration spaces also influence the focused crawling recall (Figure 8). Nowadays, focused crawler conventional just explores at quadrant I and II because the search is done just in forward direction. Quadrant III and IV have not been explored by conventional focused crawler and this study proposes a method to explore the quadrants comprehensively.

5.3. Determinants of Precision and Recall

The description of chapter 5.1 and 5.2 can be concluded that there is one major variable that influences focused crawling precision (number of query concepts) and three main variables that influence focused crawling recall (Table 1). Focused crawler has high precision when the query only relates to one main concept (Q:C = 1:1). More concepts of the query makes precision decreases exponentially Q:C = 1:|c|.

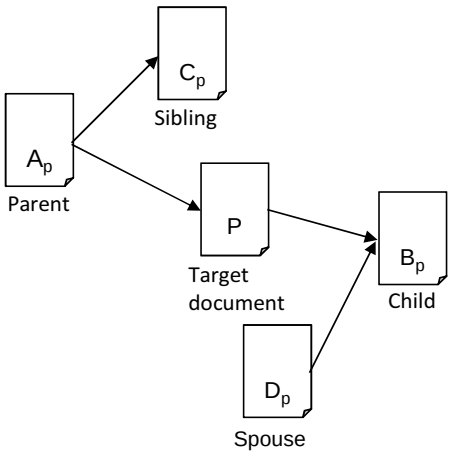
Table 1. Effect of Q (query), C (concept), L (lexicon) and K (quadrant search) on precision and recall of focused crawling system

Recall \ Precision	Low	High
High	$Q:C = 1:\|c\|$ $O_L:C = 1:\|c\|$; $C:L = 1:\|l\|$ & $P_D(K)=\{I..IV\}$	$Q:C = 1:1$ $O_L:C = 1:\|c\|$; $C:L = 1:\|l\|$ & $P_D(K)=\{I..IV\}$
Low	$Q:C = 1:\|c\|$ $O_L:C = 1:1$; $C:L = 1:1$ & $P_D(K)=\{I,II\}$	$Q:C = 1:1$ $O_L:C = 1:1$; $C:L = 1:1$ & $P_D(K)=\{I,II\}$

Focused crawling recall depends on variables of: (1) Completeness of concepts knowledge related to the topic (local ontology - O_L). More concepts of the local ontology (O_L:C = 1:|c|) will increase focused crawling recall; (2) Completeness of derivative concepts' lexicon synonyms contained in the local ontology, because of there are many lexicons which have similar meaning (synonymous). The more synonyms recognized the better increase focused crawling recall; and (3) Completeness of exploration spaces to obtain relevant documents ($P_D(K)$). As seem variable (1) and (2), the more complete search spaces can be explored, the higher focused crawling recall will be.

5.4. Focused Crawling Exploration

There are four types of neighboring documents in Web crawling search space, i.e.: parent document, child document, sibling document, and spouse documents shown in Figure 9 [14].



Ap	Parent document	$\{q \mid q \rightarrow p$ dan $q \neq p\}$
Bp	Child document	$\{q \mid p \rightarrow q$ dan $q \neq p\}$
Cp	Sibling document	$\{q \mid \exists\, r$ s.h. $r \rightarrow p$, $r \rightarrow q$ dan $q \neq p\}$
Dp	Spouse document	$\{q \mid \exists\, r$ s.h. $p \rightarrow r$, $q \rightarrow r$ dan $q \neq p\}$

Figure 9. Four kinds of document neighbors

Formula (1) and (2) are the algorithm base to obtain child and parent documents.

$$SUCC(p) = \{q \mid O_j(p) = q\} \quad (1)$$

$$SUCC^{-1}(p) = \{q \mid O_j(q) = p\} \quad (2)$$

Generally, focused crawler can only explore relevant documents in quadrant I and II. This is because focused crawler only use downloaded documents' out-links as traversal guidance. Focused crawler can explores relevant documents located in quadrant I because there exist out-links from one relevant document to others in the strong connected characteristic. Formula (3) is an algorithm to explore relevant documents in quadrant I.

$$Reachable(FC_I) = \{U(q) \mid \text{while } Score(q)=1 \text{ do } p= q, SUCC^*(p)\} \quad (3)$$

Several studies have proven the existence of relevant documents that are connected through one or more (maximum of twelve) irrelevant document. Therefore, several focused crawling methods are not cut off directly the routes through irrelevant documents but reducing the weight of encountered out-links. The farther out-link from relevant documents, the less relevance weight will be. Algorithm to explore relevant documents in quadrant II is in formula (4)

$$Reachable(FC_{II}) = \{U(q) \mid \text{while } 0<Score(q)<1 \text{ and } d(q)<12 \text{ do } p= q, SUCC^*(p)\} \quad (4)$$

In order to increase focused crawling recall, to explore relevant documents in quadrant III and IV may not be done just by utilizing the downloaded documents' out-links, but has to utilize backlinks of potential downloaded documents, too. Relevant documents in quadrant III are analogue to spouse document in Figure 9. When the downloaded relevant documents point to the same child, then the child documents can be regarded as an authority. If the child document is an authority, then all spouse documents predicted as candidates of relevant documents and must be downloaded. To clarify this, see the algorithm below to find spouse documents,

if $x=SUCC(p)$ and $x=SUCC(q)$
then
    $x \leftarrow AUTHORITY$
    $SPOUSE(p)= SUCC^{-1}(x)$

Formula (5) is an algorithm to explore relevant documents in quadrant III.

$$Reachable(FC_{III}) = \{U(q) \mid r=SUCC(p) \text{ and } s=SUCC(p); \text{ if } Score(r)=1 \text{ or } Score(s)=1 \text{ then } (SUCC^{-1})^*(p)\} \quad (5)$$

Similar to quadrant III, relevant documents of quadrant IV analogue to sibling documents in Figure 9. When the downloaded relevant documents are pointed by the same parent, then the parent documents can be regarded as a hub. If the parent document is a hub, then all sibling documents predicted as candidates of relevant documents and must be downloaded. The algorithm below is to find sibling documents,

if $y=SUCC^{-1}(p)$ and $y=SUCC^{-1}(q)$
then
    $y \leftarrow HUB$
    $SIBLING(p)= SUCC(y)$

Formula (6) is an algorithm to explore relevant documents in quadrant IV.

$$Reachable(FC_{IV}) = \{U(q) \mid p=SUCC^{-1}(r) \text{ and } p=SUCC^{-1}(s); \text{ if } Score(r)=1 \text{ and } Score(s)=1 \text{ then } SUCC^*(p)\} \quad (6)$$

Total of reachable relevant documents are,

$$Reachable(FC) = Reachable(FC_I) + Reachable(FC_{II}) + Reachable(FC_{III}) + Reachable(FC_{IV}) \quad (7)$$

Formula (7) is an algorithm to reach relevant documents which are connected to each other either directly or indirectly and connected through out-links or backlinks. Whereas to reach disconnected relevant documents, focused crawler utilizes the ontology to maximize the result.

6. Result

Experiments have been carried out crawling process with CT-FC strategy on several topics, including the topic of "algorithm". There are 1714 documents which are relevant to the topic "algorithm" in DMOZ. Many relevant URLs were taken at random as much as 1 to 80 URLs used as seed URLs and the rest are considered as target documents.

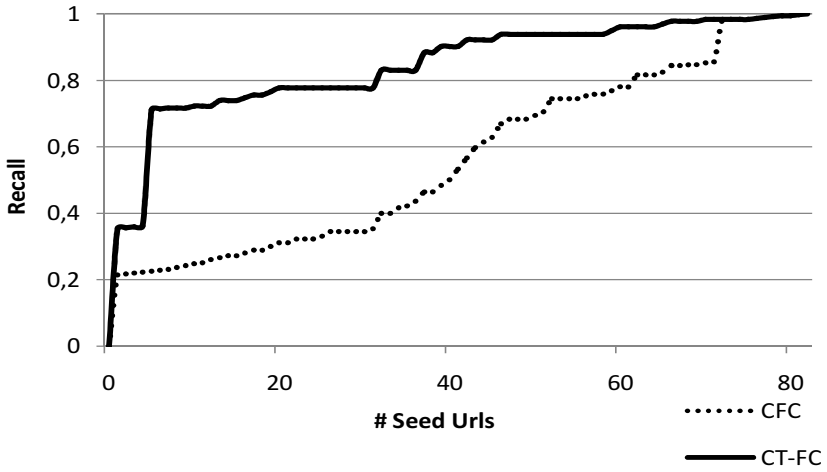


Figure 10. Average recall comparison of more Comprehensive Traversal Focused Crawler (CT-FC) and Conventional Focused Crawler (CFC)

Conventional strategy of focused crawling without using in-link information is done for comparison. Figure 10 shows the comparison of recall range average of CT-FC and conventional focused crawler. The same variation of seed URLs as well as CT-FC is given to the conventional focused crawling. CT-FC gives a significant increasing from conventional focused crawler recall. With a small seed URLs, conventional focused crawling produces recall so far just about 0.5 but with CT-FC, it quickly generates recall above 0.7 and continues increasing rapidly when the number of seed URLs added.

7. Conclusion

With the forward and backward crawling approach, focused crawler can increase the exploration capability and recall performance. With this ability, the constraints faced by conventional focused crawler associated with the Web structure characteristics can be resolved. This can be proved by the high value of crawling recall although just a small number of seed URLs is given.

This study proves the relevance support from a relevant document for sibling documents through co-citation, and to spouse documents through co-reference. Based on the result of the experiment, forward and backward crawling approach make focused crawler becomes more stable, (not sensitive to the amount and quality of seed URLs). Bibliometric concepts also supports CT-FC to have good performance, especially in precision, recall and stability.

References

[1] Chen Y. *A Novel Hybrid focused crawling algorithm to build domain-specific collections*. PhD thesis. Virginia - United States. Virginia Polytechnic Institute and State University; 2007.

[2] Maimunah S, et al. *Community Associations As A Knowledge Base To Improve Focused crawling Recall*. 5th International Conference on Information Communication Technology and Systems(ICTS). Surabaya. 2009: 225–230.
[3] Ali H. Self Ranking and Evaluation Approach for Focused Crawler Based on Multi-Agent System. *The International Arab Journal of Information Technology*. 2008; 5(2): 183–191.
[4] Qin J, Zhou Y, Chau M. *Building Domain-Specific Web Collections For Scientific Digital Libraries: A Meta-Search Enhanced Focused Crawling Method*. 4th ACM/IEEE-CS Joint Conference on Digital Libraries. Tucson AZ USA. 2004: 135–141.
[5] Heinonen O, Hatonen K, Klemettinen M. *WWW Robots and Search Engines*. Seminar on Mobile Code. Report TKO-C79. Helsinki University of Technology. Department of Computer Science. 1996.
[6] Salton G, McGill M. An Introduction to Modern Information Retrieval. McGraw-Hill. New York. 1983.
[7] Rijsbergen CJ. Information Retrieval. Butterworth. 1979.
[8] Pinkerton B. *Finding What People Want: Experiences with the WebCrawler*. Proceedings of the Second International World Wide Web Conference. 1994.
[9] Bergmark D, Lagoze C, Sbityakov A. *Focused Crawls, Tunneling and Digital Libraries*. Proc. of the 6th European Conference on Digital Libraries. Rome Italy. 2002.
[10] Kumar R, et al. *Trawlling the Web for Emerging Cyber-Communities*. Proc. of 8th International World Wide Web Conference. Toronto Canada. 1999.
[11] Kumar R, et al. *Extracting Large-Scale Knowledge Bases from the Web*. Proc. of the 25th International Conference on Very Large Data Bases Conference. Edinburgh Scotland UK. 1999a.
[12] Toyoda M, Kitsuregawa M. *Creating a Web Community Chart for Navigating Related Communities*. Proceedings of ACM Conference on Hypertext and Hypermedia. Århus Denmark. 2001: 103–112.
[13] Dean J, Henzinger MR. *Finding Related Pages in the World Wide Web.* Proceedings of the 8th International WWW Conference. Toronto, Canada. 1999: 1467–1479.
[14] Qi X, Davison BD. Knowing a web page by the company it keeps. *CIKM*. 2006: 228–237.