Master thesis

Detecting Web Spam from a Directed Graph of Web Sites

ウェブにおける有向サイトグラフからのスパム発見に関する研究

Supervisor: Professor Masaru Kitsuregawa

THE UNIVERSITY OF TOKYO

2007 February 2nd

Department of Information and Communication Engineering,
Graduate School of Information Science and Technology
The University of Tokyo

Kitsuregawa Lab: 56449 Bingshuang Han

ABSTRACT:

Link spam, which attempts to deceive link-based ranking algorithms of search engines by building densely connected structure between sites, has attracted the attention of researchers in year 2004 and 2005. It has been tightly connected with the success of commercial search engines (such as Google).

In our research, we propose a technique for detecting link spam sites in the Web. Our method detects densely connected sets of sites from a directed graph of sites based on several patterns of directed connections, such as cycles and co-citations. We discuss which patterns are useful for detecting link spam, and show results of experiments on our Japanese web archive.

The main contributions of this dissertation are outlined as follows:

- We propose a method for detecting the web spam structure based on several patterns of connections.
- We examined appropriate connection patterns and threshold for clustering the spam sites.
- We show the results of an extensive evaluation, based on 600 million sites and a manual examination of over 4000 sites.

Keywords:

Densely Connected, directed graph, Link Spam, union-find

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In this digital information age, more and more people rely on the Internet to find all kinds of information needed. Web search engines also change the way people living, shopping as well as information exchanging: without going out of the home, we can buy books, clothes, even food in the online shops. Although today's search engines can easily return millions of sites for a certain query, it is impossible for users to preview all the results. As they imply great business gains, owners of web sites expect to always be shown up on the top of result lists. This leads to the emergence of SEO (Search engine optimization) which helps web sites to acquire high ranking scores in search engines. Some examples of these techniques are: using significant titles for Web sites, giving descriptive words in the Meta tags, etc. However, there is no clear definition in the legitimation of these sites and moreover, it brings black arts such as web spamming [15] where some authors create web sites with the main purpose of misleading search engines and obtaining higher ranking than the deserved ranking .

The direct outcomes of web spam are: It plays an important role in decreasing the quality of search results returned from search engines; A large number of low-quality boosting sites are need to be crawled and indexed by search engines.

There are two categories of techniques applying in web spam[15]: one is Term Spamming, which are used to hide information from the eyes of human web users, for instance, repetition of specific terms, dumpling of a great number of unrelated terms, weaving of spam terms into contents, gluing sentence or phrases together and etc. The other one is Link Spam, which mainly belongs to sites boosting techniques. By creating specific link structure, some sites can achieve high ranking scores under the algorithms which are used to compute importance scores based on the link information.

Among these techniques in web spam, link is one of the most important and very hard to identify as well due to its dynamics and huge amount of data. Researchers said link spam will be dominant in future web spam techniques. Therefore, it is quite important to know, link spam detection and counteraction is necessary to construct a credit web society and enhance the quality of results returned by search engines.

1.2 Objective and Scope of Study

Problem definition:

- Web spam is a kind of technique with the sole purpose to mislead search engines in giving undeserved high ranking scores to some intended sites (usually, these sites are called target sites.)

- Link spam is one kind of Web spam, which is used to cheat the connectivity-based ranking algorithms in order to increase the ranking of target sites.

- Special link structure among web sites helps to detect web spam. So, how to extract link structure becomes our significant destination.

Main research content:

Our research focuses on densely connected link structure analysis, as spammers always boost a huge amount web sites interconnected each other to increase the ranking score of target sites. According to the limit cost in hardware configuration and human managements, spammers will take the advantages of boosted sites as much as they can. It is not hard to see, as a marked characteristic in the result, these boosted sites should be highly density connected each other. For a macro view of neighborhood link structure, we can conclude several link structure patterns. Use these patterns to match the whole network, and then record the web sites in the shape of clustering (These clusters are called link spam farm, introduced in Chapter 2.3).

In this thesis, besides the advanced analysis on link structures, we present a technique for detecting web spam from a densely connected directed graph of sites. The outline of this work is

1) Construct link structure patterns used in our experiment. As three-node-pattern is the smallest unit in mining linkage, so our objective point mainly focuses on here.

2) Matching the whole network with link structure patterns in 1), record all the sites that satisfied these conditions, and then do clustering, with the destination of detecting spam farm.

3) Do manual examination of the contents of web sites recorded in 2) to conclude which pattern is the most appropriate for detection link sites. The do sample test using labeled non-spam sites, check the distribution of these sites in each clustering results in 2) to figure out which pattern in 1) is the best one in eliminating non-spam sites.

4) From a macro view, do the analysis based on combining all the results of experiments, and ascertain the suitable threshold value in each experiment in detail.

5) Do visualization of spam farm recorded and analysis.

6) Make a conclusion and discuss the future work.

1.3 Organization of the Thesis

The thesis is organized as follows: the background and related work is introduced in Chapter 2. Preparatory work is introduced in Chapter 3.Web extraction patterns are explained in Chapter 4. The experimental results and analysis of the results are shown in Chapter 5, and we conclude the paper in Chapter 6.

Chapter 2

Preliminaries

2.1 PageRank

Usually, current search engines combine several algorithms to calculate the ranking score of sites. One of the most famous of them is PageRank, in which, the authority reminds the notion of citation in the scientific literature, particularly, the authority of a site p depends on the number of incoming hyperlinks (num of citations) and on the authority of the site q which cites p with a forward link [9].

We adopt graph $G=(v,\varepsilon)$ to express the web of interlinked hypertext documents, while v denotes a web site (a node) and ε shows the link information <p,q> between web sites (the edge between nodes).

The core algorithm in PageRank is:[16]

$$p_0 = d \sum_{q \in q[p]} \frac{p_i}{h_q} + (1-d)$$

Where, p_0 is the PageRank score of site p. d ∈ (1, 0) is a dumping factor , got by experiments to guarantee convergence and h_q is the out degree of q, which means the number of hyperlinks out-going from q.

The equivalent matrix notation is following:

$$p = d \cdot T' \cdot p + (1-d)\frac{1}{\|V\|} I_{\|V\|}$$

Where T' is transposed transition matrix, T = $(T_{ij})_{n\times n}$ defined as

$$T_{i.j} = \begin{cases} 1/h_q & if\ (i,j) \in \varepsilon \\ 0 & otherwise \end{cases}$$

This function describes a transition state that the browser doing a random visit on the web sites, starting at a random site p, and at each step follow one of h_q links on that site each with $1/h_q$ probability. It is easy to draw a conclusion: 3 elements will be helpful to enhance the PageRank scores of web sites:

- The number of out-links: simple measure of welcomeness.
- Whether the out-links from recommended sites or not: absolute measure of welcomeness
- The link number of out-link sites: the possibility of target site to be chosen.

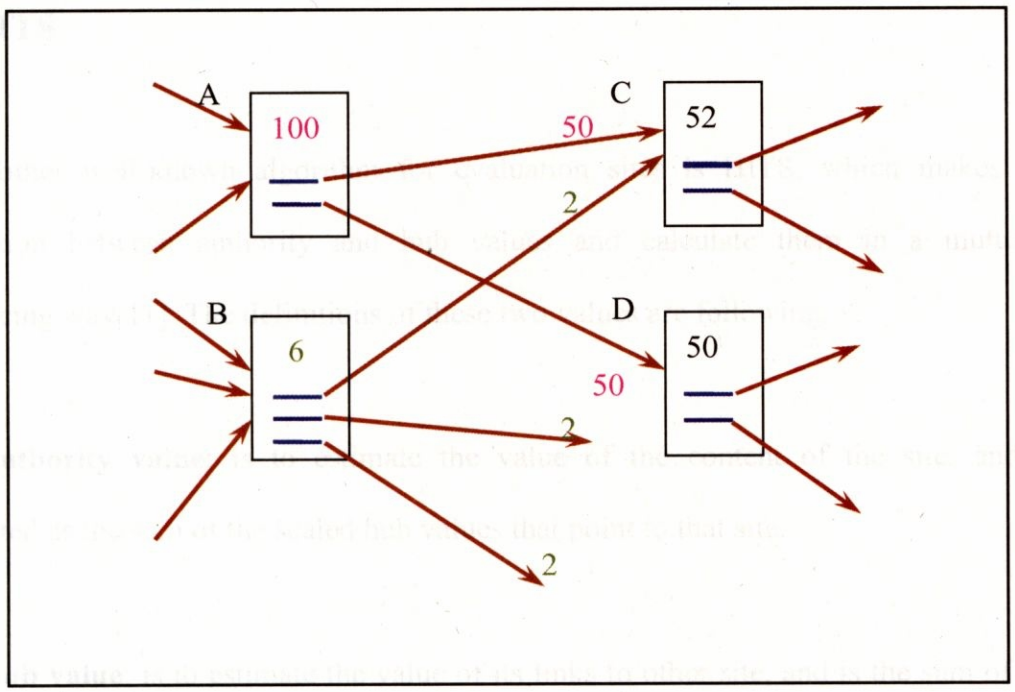


Figure 2.1: Simplified PageRank calculation

Figure 2.1 is the simplified case of link structure of four nodes A, B, C, and D. Node A and B are back-links of node C while Node B is back-link of Node A. The arrowheads show how PageRank scores of Node A and B contribute Node C and D.

Since PageRank score doesn't contain any valuable information about the quality of the contents of web sites except linkage information. Apparently, it became a natural attacking target for spammers.

Therefore, spammers usually set up a large group of web sites with densely interconnected link structures for increasing the ranking of several target sites.

2.2 HITS

Another well-known algorithm for evaluation sites is HITS, which makes the distinction between authority and hub values and calculate them in a mutually reinforcing way[11]. The definitions of these two values are following:

Authority value: is to estimate the value of the content of the site, and is computed as the sum of the scaled hub values that point to that site.

Hub value: is to estimate the value of its links to other site, and is the sum of the scaled authority values of the sites it points to.

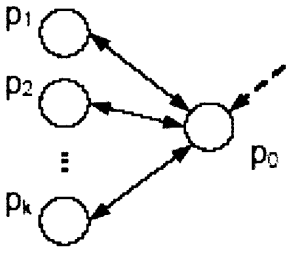
The intuition is that, an important hub points to many important authorities and an important authority links to many important hubs.

HITS, likes PageRank, and is one algorithm based purely on the link structure of web sites. However, the search engine using HITS algorithm will give the results with high ranking score in both hug sites and authority sites. Some web directory services allow free link registration, and spammers often submit links to these web directory services to boost authority score.

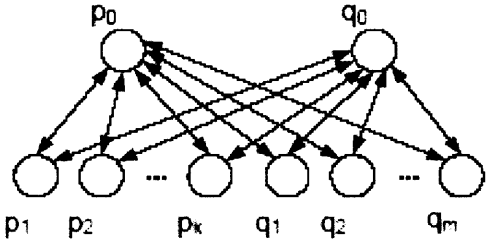
2.3 Link farm patterns

- Link spam farm patterns: As mentioned in Hector's "Link Spam Alliances"[16].

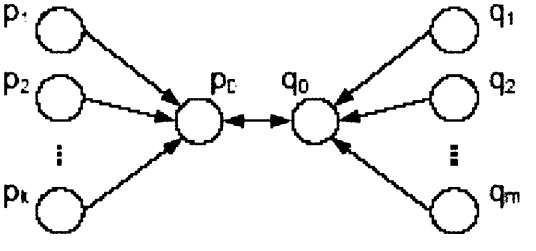
Link spam is to create link structures of interconnected sites artificially to indirectly affect search engines into boosting higher-than-deserved ranking scores. In these link structure patterns, except target site, all the other sites are called boost sites, and these group are called link spam farms. For instance:



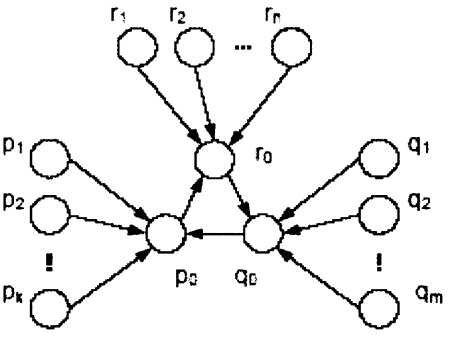
(a) A simple spam farm with one target site



(b) Two spam farms



(c) Two spam farms with interlinked target sites



(d) Three spam farms

Figure2.2: Possible link farm patterns

In figure 2.2 (a) (b) (c) (d), p_0, q_0, r_0 belong to target sites while $p_1, p_2, \cdots, p_k$, $q_1, q_2, \cdots, q_m$ and $r_1.r_2, \cdots, r_m$ belong to boosting sites.

2.4 Related work

While the term "spam" appeared as early as in year 1996, link spam, as one way of web spam, has acquired a highlighted position by year 2004. From its appearance, link spam attracted the attention of researchers in database, IR and Web. However, the research on the particular issues of link spam, i.e., huge amount of data, on-line processing, is still at its starting and most works focus on algorithm development to identify link spam.

Especially, in these several field:

- Academic co-citation and co-reference analysis

- Link-based graph analysis
- Link structure based clustering algorithm
- Hubs and Authorities patterns

Hector Geocia-Monila presented the "Link Spam Alliances" to give a detailed analysis for how spam farm can optimize web sites ranking by interconnecting each other and their results also shows the optimal structures of spam farm and quantify the potential gains in ranking score[16].

Davison showed the idea about recognizing and eliminating nepotistic links and demonstrated recognition of such links with high accuracy automatically is potential [10].

Broder and Bharat analyzed large amount of web sites and observed that the in-degree and out-degree should follow Zipfian's law and found that "artificially" generated link farms are the outliers in the distribution [7].

Fetterly analyzed the distribution of many web site features over 429 million sites. They found the sites generated automatically are quite different from the sites authored by a human; moreover, they described several properties that help to indicate web spam sites [13,14].

Gyongyi et al. described a new algorithm, Trustrank, to combat Web Spam [17]. The basic idea is that good sites always link to good sites and seldom point to bad ones.

They first selected seed sites, and then these seed sites propagate trust weights along the hyperlinks. However, the manual selection of trusted sites creates a perceptive bias as unknown and remote websites become less visible. Wu et al proposed Topical TrustRank in 2006; they made a supplement by arrange the site the topic biased TrustRank. [23]

Wu et al. introduced one algorithm to detect link farms [22]. They first chose the seed set (spam sites) based on common link structure between incoming and outgoing links of the web sites. Then they expanded the seed set with the nodes that have too many outgoing links to the original seed set.

Chapter 3

Preparatory work

3.1 Visualization of neighborhood link structure

In order to confirm the existence of link spam farm-based web spam in the top results given by the current search engines. We first did a simple investigate on visualizing the neighborhood link structure based on the top 100 result of keyword Japanese "傘". We wish to exam how real link spam farm appears in the current network and attempt to establish link pattern model which could help to identify link spam sites.

The concrete steps are shown as follows:

1 Put the keyword as a query topic in search engine; here we put Japanese character 傘 in Google. Make the top 100 web sites from all the result as target sites into the seed sites list.

2. For each site p in target list, claw its in-links and out-links and preserved these link information into edge set. Base on their distance to target sites, we make the

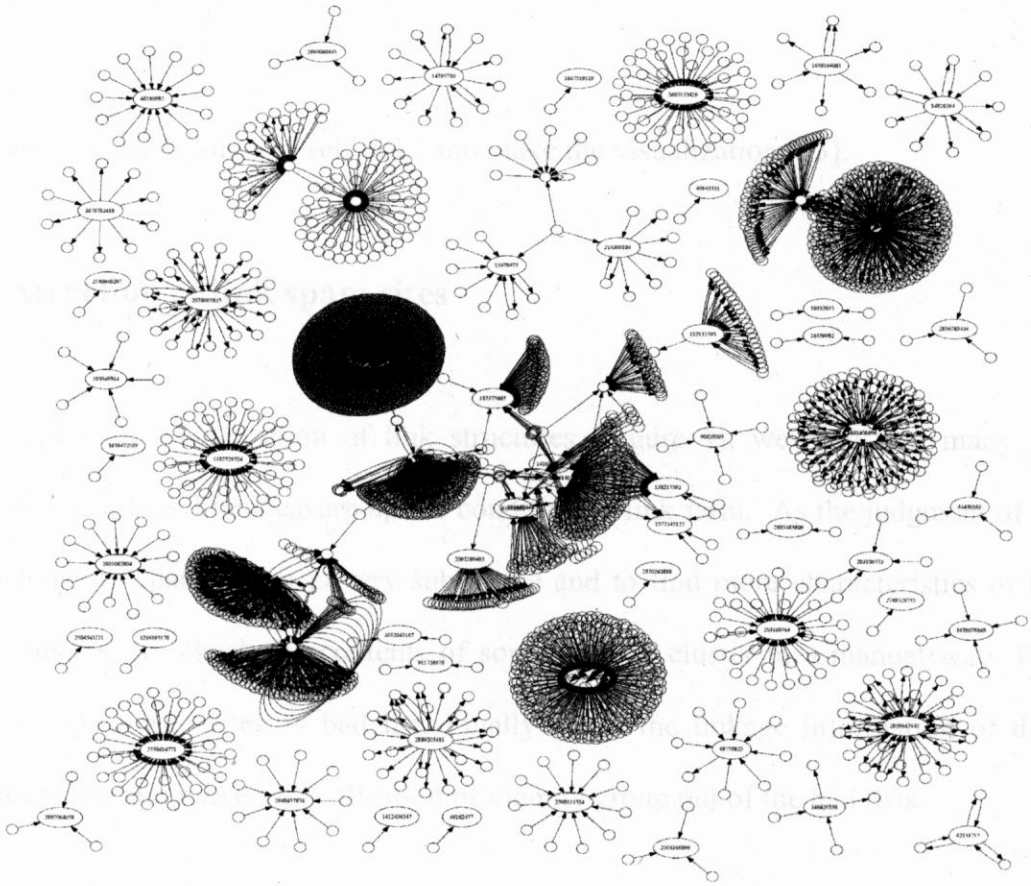


Figure 3.1: The graph of neighborhood link structure based on the top 100 result of keyword "傘"

setting of distance Degree.

E.g. edge <A, B> distance Degree of A to B is 1; <A, C>, <C, B> distance Degree of A to B is 2.

3. Remove multi-link and self links and make the visualization [13].

3.2 Extraction of link spam sites

According to visualization of link structures (Figure 1), we may find many link clusters which look like measure up the condition of link farm. As the judgment of one site belongs to spam or not is very subjective and to find more characteristics of link spam clusters, we check the contents of some sites in cluster in a manual way. Then keep the spam web sites as bad list. Finally select the linkage information of these reference sites and make a visualization of vicinity group out of the bad lists.

Figure 3.1 showed the graph of neighborhood link structure of these 100 sites. Particularly the ellipses are use to presented the target sites, and the number inside is their Unique ID used by our data sets. Each circle is instead of a web site of degree 1 or 2. And arrowheads show the relationship of pointing direction in web sites.

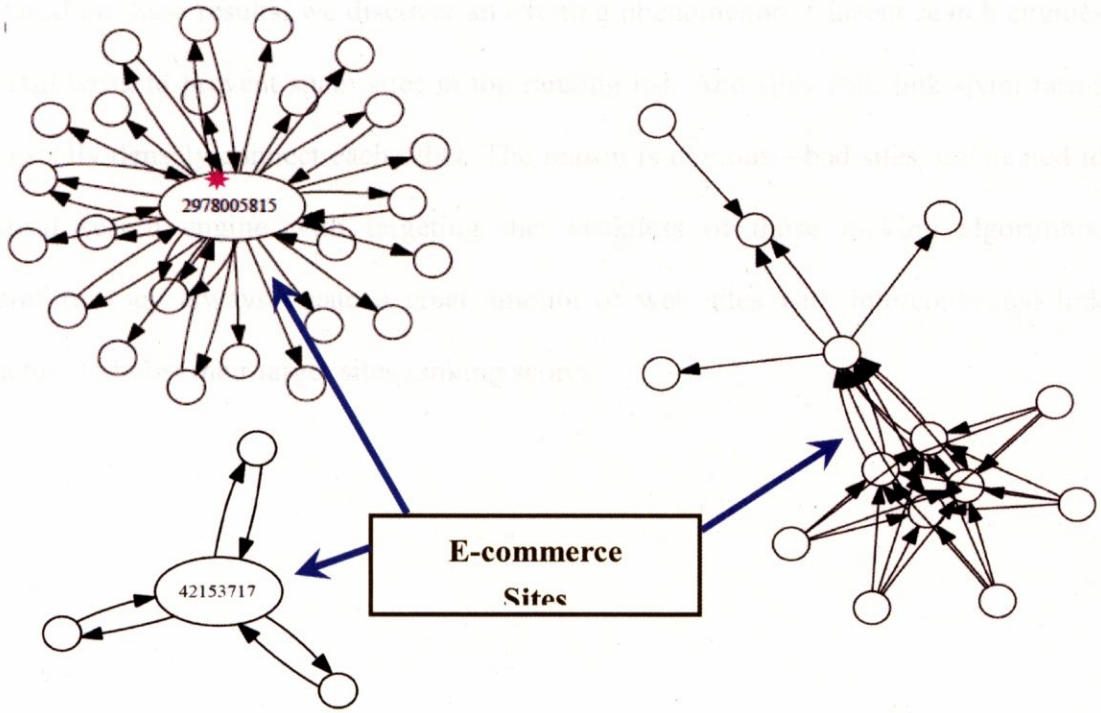


Figure 3.2: The visualization of part spam clusters in Figure 1

In order to display the result distinctly, we defined the classification of the subjects in [21]:

- Non-Spam: Sites offer useful information, including personal sites, corporation sites and sites of public institutions.
- Link Directory & Sales Promotion: the content of sites just consists of link information and the advertisement about product, such as discount information and etc.
- Pornographic sites
- Unsure: unknown language

Based on these results, we discover an exciting phenomenon: Current search engines are still weak to prevent spam sites in top ranking list. And sites with link spam farms are usually densely connect each other. The reason is obvious—bad sites are created to mislead search engine with targeting the weakness of those ranking algorithms. Therefore, they always create a great amount of web sites with interconnected link structure to boost their target sites ranking scores.

Chapter 4

Web Spam Extraction Methods and Algorithms

4.1 Web model

We adapted the web model as directed graph G (V, ع) consisting of:

V, a set of web sites (vertices)

ع, a set of links(edges) between web sites.

Besides, we use other symbols following:

<p, q> : represent a link from site p to site q.

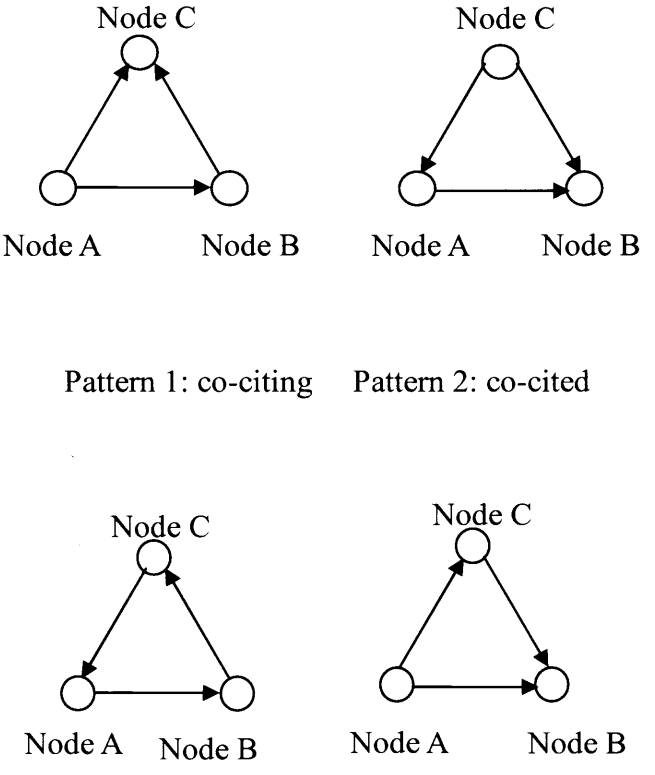
In (p): a set of in-link sites of site p.

Out (p): a set of out-link sites of site p.

For the sake of simplicity but without loss of generality, we collapse multiple links between two nodes into a single link, and also remove self links from .ع

4.2 Patterns extraction

Let us consider the link connection among three nodes, the smallest link farm alliance unit in the network. Suppose three are nodes A, B, C and between each two nodes there exists one linkage. The link direction between nodes A and B is determined, i.e., from A to B. Link directions between node C and nodes A, B are unknown. Note that there are four possible link structures for these three nodes. We give the definition of each pattern as follows (shown in Figure 1):



Pattern 1: co-citing Pattern 2: co-cited

Pattern 3: circle Pattern 4: support

Figure 4.1: The definition of 4 patterns

Pattern 1: Co-citing. Both node A and B are co-citing to node C.

Pattern 2: Co-cited. Both node A and B are co-cited from node C.

Pattern 3: Circle. Node A is co-cited from node C and node B is co-citing to node C.

Pattern 4: Support. Node A is co-citing to node C and node B is co-cited from node C.

The complete algorithm is shown in Figure 4.2. In the implementation, we extract the incoming sites and outgoing sites of both nodes A and B, and denote them as In (A) (incoming sites list of A), Out (A) (outgoing sites list of A), In (B) (incoming sites list of B) and Out (B) (outgoing sites list of B). To extract the four patterns, we just need to compare the corresponding sets to obtain the common shared nodes in these sets.

Concretely, in each pattern:

- Co-citing pattern: Compare Out (A) and Out (B) two sets and output shared node C.
- Co-cited pattern: Compare Out (A) and Out (B) two sets and output shared node C.
- Circle pattern: Compare In (A) and Out (B) two sets and output shared node C.
- Support pattern: Compare Out (A) and In (B) two sets and output shared node C.

The computation complexity of this algorithm is O (d * |E|) where d is maximum degree of nodes, and |E| is the number of edges in the Web graph.

In order to have a better understanding of shared node C size distribution for each pattern, we collect the size information of shared node C and made them into graphs and tables (See Figure 5.4). This distribution information helps us to know the main portion

Pattern extraction method

Input:

P Node

Np.in The in-link number of P

Np.out The out-link number of P

E (A, B) The edge from A to B

- Cluster node C (E.g Pattern 1)

  For each edge in G (P, E)

  Edge = (A->B) ($A, B \in P$)

  Out (B) =out-links (B)

  Out (A) =out-links (A)

  Num =number of nodes shared between

(Out (B), Out (A))

Output:

(A, B, num)

Figure 4.2: The algorithm for pattern extraction of node C

(In the case of Pattern 1)

distribution range of node A and B which play an important role to decide the threshold value for union-find clustering in next step, namely, we want to do investigate on the main portion nodes which satisfy each pattern.

4.3 Clustering based on union-find method

In the previous step, we obtained the results of node A and B's common sharing node C in each pattern. Moreover, we deduced the distribution of the number of nodes per cluster in each pattern with statistical information. These relationship charts will be helpful to determine the threshold N in the following step. (See Chapter 5.3)

In order to create the vicinal density connected nodes graph, we employ clustering in nodes set with respect to the number of nodes being shared. The method used to merge nodes set is union-find algorithm which has high efficiency in performing Find and Union operations on disjoint-set data structure.

In practice, if the number of the node C shared between A and B is more than N (threshold), we do union operation to merge the nodes sets or cluster sets into final cluster sets in each pattern.

4.4 Sample examination of non-spam sites

In this step, let us evaluate the result in an important aspect on how non-spam sites are included in the result cluster sets. Since it is impossible to check all the sites' content,

here we apply some sample test. First, we manually collect some non-spam sites, which have a great number of incoming and outgoing links, from the nodes in original dataset and make them into a white-site list. Then, we identify how many labeled non-spam sites exist in cluster sets for each pattern with respect to different thresholds. By this way, we can determine the appropriate threshold *N* and the appropriate pattern for spam detection.

4.5 Visualization of link structure

In order to display the neighbor hood link structure of link spam sites intuitionisticly. We did the visualization of partial sites in cluster units.

Method of link information extraction of sites shows as following:

- Cluster (nodes) selecting: Manually checked the contents of web sites which are included in the result of step 4.3 in cluster units. Once confirm more than 80 % of sites in this cluster are spam sites, we record these nodes ID.
- Link information extraction: Extracted all the linkage information of nodes which are related with the recorded nodes.
- Nodes display: If the In (P) >10, we keep the nodes ID, and display the node in ellipse shape (target site), else ignore the nodes ID, and put small circle to present sites instead.