The Link Structure of Language Communities and its Implication for Language-specific Crawling

Rizza Camus Caminero
Language Observatory
Nagaoka University of Technology
Nagaoka, Niigata, Japan
rhyze.caminero@gmail.com

Yoshiki Mikami
Language Observatory
Nagaoka University of Technology
Nagaoka, Niigata, Japan
mikami@kjs.nagaokaut.ac.jp

Abstract

Since its inception, the World Wide Web (WWW) has exponentially grown to shelter billions of monolingual and multilingual web pages that can be navigated through hyperlinks. Its structural properties provide useful information in presenting the socio-linguistic properties of the web. In this study, about 26 million web pages under the South East Asian country code top-level-domains (ccTLDs) are analyzed, several language communities are identified, and the graph structure of these communities are analyzed. The distance between language communities are calculated by a distance metrics based on the number of outgoing links between web pages. Intermediary languages are identified by graph analysis. By creating a language subgraph, the size and diameter of its strongly-connected components are derived, as these values are useful parameters for language-specific crawling. Performing a link structure analysis of the web pages can be a useful tool for socio-linguistic and technical research purposes.

1 Introduction

The World Wide Web contains interlinked documents, called web pages, navigable by hyperlinks. Since its creation, it has grown to contain billions of web pages and several billion hyperlinks. These web pages are created by millions of people from all parts of the world. Each web page contains a large amount of information that can be shared and disseminated to people with Internet access. Authors and creators of a web page come from different backgrounds, different cultures, and different languages. Thus, a web page is a resource of multilingual content, and a fertile source for socio-linguistic analysis.

1.1 Web Crawling

While search-engines are important means of accessing the web for most users, automated systems of retrieving information from the web have been developed. These systems are called web crawlers, where a software agent is given a list of pages to visit. As the crawler visits these pages, it follows their outgoing links and adds these to the list of pages to visit. Each page is visited recursively according to some sets of policies (e.g. the type of pages to retrieve, direction, the maximum depth from the URL (uniform resource locator), etc.). The result of this crawl is a vast amount of data, most of which may be irrelevant to certain individuals. Thus, a focused-crawling approach was implemented by some systems to limit the search to only a subset of the web.

Focused crawlers rely on classifiers to work effectively. Language-specific crawlers, for example, need a very good language identification module that properly identifies the language of the web pages. General crawlers can be extended to include focused-crawling capabilities by incorporating the classifiers. Another important requirement to efficiently crawl the desired domain is the list of initial web pages, called seed URLs. Each of these URLs will be enqueued into a list. The agent visits each URL in the list. Since the crawler recursively visits

the outgoing links of each URL, it is possible that the seed URL is an outgoing link of another seed URL, or an outgoing link of an outgoing link of another seed URL. Listing these URLs as seed URLs will just waste the crawler's time in visiting them, when they have already been visited. If several URLs can be reached from just one URL, the seed URL list size will decrease and the crawler will be more efficient. However, the maximum distance between these web pages must also be considered, since the crawler can have a policy to stop the crawling after reaching a certain depth.

1.2 The Web as Graph

A graph consists of a set of nodes, denoted by *V* and a set of edges, denoted by *E*. Each edge is a pair of nodes *(u, v)* representing a connection between *u* and *v*. A path between two nodes is a sequence of edges that is passed through from one node to reach the other node.

In a directed graph, each edge is an ordered pair of nodes. A path from *u* to *v* does not imply a path from *v* to *u*. The distance between any two nodes is the number of edges in a shortest path connecting them. The diameter of the graph is the maximum distance between any two nodes.

In an undirected graph, each edge is an unordered pair of nodes. There is an edge between *u* and *v* if there is a link between *u* and *v*, regardless of which node is the source of the link.

A strongly-connected component (*SCC*) of a directed graph is a set of nodes such that for any pair of nodes *u* and *v* in the set, there is a path from *u* to *v*. The strongly-connected components of a graph consist of disjoint sets of nodes. A *subgraph* is a graph whose nodes and edges are subsets of another graph.

With the interlinked nature of the web, it can be represented as a graph, with the web pages as the nodes and the edges as the hyperlinks between the pages.

1.3 Languages

Languages are expressions of individuals and groups of individuals, essential to all form of communication. It is a fundamental medium of expressing one's self whether in spoken or written form. Ehtnologue (2005) lists 6,912 known living languages in the world. Only a small portion of these languages can be found in the web today.

A language community in the web is the group of web pages written in a language. Major language communities discovered in each country indicates the dominant language of the country's web space. How one language community is related to another language community can be shown by analyzing the hyperlinks between them. Thus, a language graph can be created with the language communities as nodes, and the links between the language communities as edges.

2 Previous Studies

One of the earliest web survey in Asia (Ciolek, 1998) presented statistical data of the Asian web space by using the Altavista WWW search engine in gathering its data. In 2001, he wrote a paper presenting the trends in the volume of hyperlinks connecting websites in 10 Eash Asian countries.

Several studies have also been done regarding the representation of the web as a graph. Kumar et al. (2000) showed that a graph can be induced by the hyperlinks between pages. Measures on the connected component sizes and diameter were presented to show the high-level structure of the web. Broder et al. (2000) did experiments on the web on a larger scale and showed the web's macroscopic structure consisting of the SCC, IN, OUT, and TENDRILS. Balakrishnan and Deo (2006) observed that the number of edges grow superlinearly with the number of nodes, showing the degree distributions and diameter. Petricek et al. (2006) used web graph structural metrics to measure properties such as the distance between two random pages and interconnectedness of e-government websites. Bharat et al. (2001) studied the macro-structure of the web, showing the linkage between web sites by creating the "hostgraph", with the nodes representing the hosts and the edges as the count of hyperlinks between pages on the corresponding hosts.

Chakrabarti et al. (1999) proposed a new approach to topic-specific web resource discovery by creating a focused crawler that selectively retrieves web pages relevant to a pre-defined set of topics. Stamatakis et al. (2003) created CROSSMARC, a focused web crawler that collects domain-specific web pages. Deligenti et al. (2000) presented a focused crawling algorithm that builds a model for the context within which relevant pages to a topic occur on the web. Pingali et al. (2006) created an Indian search engine with a language identification

module that returns a language only if the number of words in a web page are above a given threshold value. The web pages were transliterated first into UTF-8 encoding. Tamura et al. (2007) presented a simulation study of language specific crawling and proposed a method for selectively collecting web pages written in a specific language by doing a linguistic graph analysis of real web data, and then transforming them into variation of link selection strategies.

Despite several studies on web graph and language-specific crawling, no study has been done showing the "language graph". Herring et al. (2007) showed a study on language networks of a selected web community, LiveJournal, but not on the web as a whole.

3 Scope and Objectives of the Study

This research was conducted on the 10 ccTLDs of the South East Asian countries. This paper, however, will only show the results for the Indonesian domain (.id).

This research aims to show the socio-linguistic properties of the language communities in each country at the macroscopic level. The web page distribution for each language community in a given ccTLD and its most frequently linked to languages are shown. The distance is also computed and the language graph is illustrated.

This research also aims to show the graph properties of some Filipino language communities and its implication for crawling. Graph properties like the SCC size and the diameter will be presented to show these characteristics in a subset of the web.

Finally, this research demonstrates the usefulness of graph analysis approaches.

4 Methodology

This study was conducted by performing a series of steps from the collection of data to the presentation of the results through images.

4.1 UbiCrawler

UbiCrawler (Boldi et al., 2004) was used to download the web pages under the Asian ccTLDs. These pages were downloaded primarily for the purpose of assessing the usage level of each language in cyberspace, one of the objectives of the Language Observatory Project (LOP)[1]. The crawl was started on July 5, 2006, running for 14 days. 107,168,733 web pages were collected from 43 ccTLDs in Asia. Each page contains several information such as the character set and outgoing links. For this study, the URL of a web page and the URL of its outgoing links were used. Although there are many web pages of Asia that can be found in generic domains, they were not included in this survey to limit the volume of the crawl data.

4.2 Language Identification

A language identification module (LIM) developed for LOP based on the n-gram method (Suzuki et al., 2002) was used to identify on what language a page is written in. This method first creates byte sequences for each training text of a language. It then checks the byte sequences of the web pages that match the byte sequences of the training texts. The language having the highest matching rate is considered as the language of the web page. The language identification module used the parallel corpus of the Universal Declaration of Human Rights (UDHR), translated into several languages. After crawling, LIM was executed to identify the languages of each downloaded web page. The identification result was stored in a LIM result file that contains the URL, the language, and matching rate, among others. In this study, the issues regarding the accuracy of LIM will not be discussed.

4.3 Web Page Analysis

For this study, the web pages of the 10 South East Asian ccTLDs were selected for analysis. There were 26,196,823 web pages downloaded under these ccTLDs. The web pages for each country were grouped by languages. The list of languages was narrowed down to 20 based on the number of pages, arranged from highest to lowest.

The link structure can be analyzed by traversing the outgoing links of each web page. For each web page, its outgoing links are retrieved. For each outgoing link, the LIM result file is checked for its language. The number of outgoing links in each language is counted. If the URL of the outgoing link is not on the file, it wasn't downloaded. Therefore, the language of the ougoing link is unidentified. This is usually the case of outgoing links un-

[1] http://www.language-observatory.org

der the generic TLDs (e.g., .com, .org, .gov, etc.) and non-Asian ccTLDs.

4.4 Language Graph

There is a link between two languages if there is at least one outgoing link from a web page in one language to the other language. The language graph is created through contraction procedure, where all edges linking the same language page are contracted.

4.5 Language Adjacency Matrix

Based on the number of web pages in a language and the number of outgoing links from one language to another, the language adjacency table N for each country is created. The row and column headers are the same – the top 20 languages based on the number of web pages. The value N_{ij} is the number of outgoing links from language i to language j.

The language adjacency matrix P contains the ratio of the number of outgoing links and the total number of outgoing links as can be found in the language adjacency table. Each cell value, P_{ij} is the probability that a web page in a language i has an outgoing link to language j.

$$P_{ij} = N_{ij} / \sum_k N_{ik}$$

A link from language i to language j is not necessarily accompanied by a link from language j to i. Even if there is a link, the number of outgoing links is not equal. To show the relationship between two languages based on the link structure, the language distance is computed.

4.6 Distance between Languages

The distance between two languages measures their level of connectedness. It is the relationship between the number of outgoing links from language i to language j and vice versa. The distance is computed as the ratio of the number of outgoing links between two languages and the total number of outgoing links of the two languages.

The distance between language i and language j, D_{ij} is, $\mathrm{D_{ij}} = \sqrt{1/(\mathrm{R_{ij}} + \alpha)} - \beta$ where R_{ij} is the language link ratio is defined as,

$$R_{ij} = (N_{ij} + N_{ji}) / \left(\sum_k N_{ik} + \sum_k N_{jk} \right) \quad \text{for } (i \neq j)$$

$$R_{ij} = 1 \quad \text{for } (i = j)$$

where α is an adjusting parameter introduced to avoid division-by-zero, which may happen when $R_{ij}=0$, i.e. no links between two languages. We set α=0.0001. Thus, the maximum distance between languages becomes 99. β is another adjusting parameter to make D_{ij} as a distance metrics, and we set β=1. Assumption behind this definition is based on commonly-observed rules in our world. It is widely observed that interaction between two objects is proportional to the inverse square of distance between two objects. The number of web links between two language communities is considered as a kind of interaction. Languages with no links between them have a distance of 99, and the distance of a language to itself is 0.

Based on this distance metrics, the macroscopic language graph is created. A distance limit of 15 is used to clearly show which languages are closely-related by their link structure.

4.7 Intermediary Languages

Considering the direction of the outgoing links, the possibility that language i will link to language j may be lesser than the possibility that language i will link to language j by passing through an intermediary language k, such that $P_{ik}P_{kj} > P_{ij.}$ The intermediary language is identified, as this would mean that there are better ways to reach another language from one language.

4.8 Graph Analysis using JGraphT

JGraphT is a free Java graph library that provides graph objects and algorithms. The library provides classes that calculates and returns the strongly-connected components subgraphs. To compute the distance between nodes, several graph searching algorithms are available, one of which is the Dijikstra algorithm that computes for the shortest path. A utility to export the graph into a format readable by most graph visualization tools is also available. A graph file, written in the DOT language (a plain text graph description language) was created, containing the nodes and edges of language pages. From this, the strongly-connected components and its properties (i.e. size, diameter) were determined.

4.9 GraphVis: Visualization Tool

GraphViz is open-source graph visualization software that takes descriptions of graphs in a simple text language and makes diagrams in

several formats, including images. The neato layout, which makes "spring model" layouts, was used to visualize the distance between languages. However, the calculated distance cannot be drawn exactly, and this visualization is only two-dimensional. So, the images are distorted and do not illustrate the exact distance, only an approximation.

5 Results

This section shows some results on the Indonesian domain.

5.1 Link Structure

Indonesia is a country with one of the biggest language diversity in the world. According to Ethnologue[2], 742 languages are spoken in the country. But, the LIM results show that only five of these languages are listed in the top 10 languages in the country, i.e., Javanese, Indonesian, Malay, Sundanese, and Madurese.

No.	Language	# of Pages	# of Outgoing Links
1	Javanese (jav)	797,300 (28.01%)	33,411,032 (27.20%)
2	English (eng)	743,457 (26.11%)	16,645,014 (13.55%)
3	Indonesian (ind)	516,528 (18.14%)	20,783,793 (16.92%)
4	Thai (tha)	218,453 (7.67%)	8,952,101 (7.29%)
5	Malay (mly)	197,535 (3.47%)	4,990,402 (4.06%)
6	Sundanese (sun)	98,835 (3.47%)	5,349,194 (4.35%)
7	Luxemburg[3] (ltz)	43,376 (1.52%)	2,307,602 (1.88%)
8	Occitan[4] (lnc)	27,663 (0.97%)	351,318 (0.29%)
9	Madurese (mad)	22,121 (0.78%)	777,903 (0.63%)
10	Tatar (tat)	20,709 (0.73%)	3,334,651 (2.71%)
	Others	160,917 (5.65%)	25,930,885 (21.11%)
	Total	2,846,894	122,833,898

Table 1. LIM result for Indonesian Domain

Javanese is the most popular language in the web space of Indonesia, and constitutes 28% of the total number of web pages. It is followed by English and Indonesian. Although Indonesian is an official language of the country, it is ranked third.

English, a major business language of Indonesia, has the second largest number of web pages in the domain. But, Indonesian occupies the second rank in the number of outgoing links.

No.	Language	Languages Linked to (# of outgoing links)
1	Javanese (jav)	Javanese (22,641,844)
		Indonesian (3,761,205)
2	English (eng)	English (11,036,726)
		Javanese (1,443,054)
3	Indonesian (ind)	Indonesian (12,530,636)
		Javanese (4,168,734)
4	Thai (tha)	Thai (4,367,895)
		English (1,775,132)
5	Malay (mly)	Malay (1,944,430)
		Indonesian (1,516,778)
6	Sundanese (sun)	Javanese (2,004,316)
		Sundanese (1,641,623)
7	Luxemburg (ltz)	Luxemburg (1,128,342)
		Javanese (393,524)
8	Occitan (lnc)	Occitan (119,734)
		English (83,380)
9	Madurese (mad)	Madurese (387,585)
		Javanese (93,837)
10	Tatar (tat)	Javanese (1,802,601)
		Thai (397,988)

Table 2. Language Link for Indonesian Domain

The table above only shows the top 10 languages. Among these, 8 languages are most frequently linked to the same language. The two other languages are Sundanese and Tatar, both mostly linked to Javanese.

The language graph below shows the languages as the nodes and the edges representing the distance between languages. In the figure above, the six languages of Indonesia are found to be closely-connected to each other.

[2] Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International.
[3] Luxembourgish
[4] Occitan Languidocien

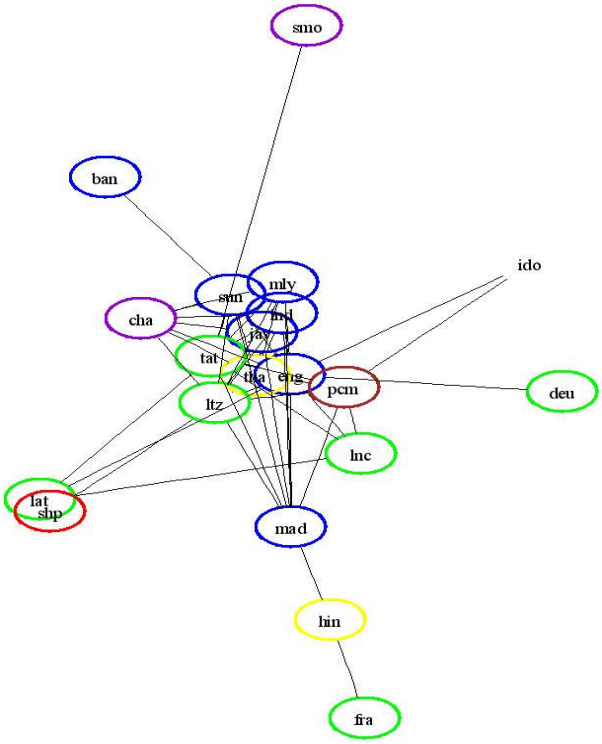

Figure 1. Language Graph of Indonesia

5.2 Intermediary Languages

For languages with only a few outgoing links between them, there exists a language that acts as its intermediary, that makes access between the two languages more convenient if passed through.

No.	Intermediary Language	Frequency	Percentage
1	English	89	17.73%
2	Javanese	43	8.57%
3	Tatar	42	8.37%
4	Thai	40	7.97%
5	Madurese	38	7.57%
6	Indonesian	34	6.77%
7	Latin	34	6.77%
8	Sundanese	29	5.78%
9	Malay	28	5.58%
10	Luxemburg	24	4.78%
	Others (top11-20)	122	24.30%
	Total	502	100.00%

Table 3. Intermediary Languages of Indonesia

The above table shows the number of language pairs having the given language as its intermediary language. English has the highest frequency as an intermediary language. However, it is likely that several pages were misidentified by LIM. The second, Javanese is not surprising since it is a major language of Indonesia. The table below shows selected language pairs, where one of its intermediary languages is Javanese.

Language *i*	Language *j*	P_{ij}	$P_{ik}*P_{kj}$ [5]
Tatar	Indonesian	0.01753	0.06085
Tatar	Luxemburg	0.00210	0.01193
Samoan	Balinese	0.00000	0.00052

Table 4. Selected Languages of Indonesia in which it's Intermediary Language is Javanese

5.3 Graph Properties

This section discusses the size distribution of the SCCs and the diameter of the Filipino language community in Indonesia.

SCC size

The SCCs of a graph are those sets of nodes such that for every node, there is a path to all the other nodes. The size of each SCC refers to the number of nodes it contains. Distribution of the sizes of SCCs gives a good understanding of the graph structure of the web, and has important implications for crawling. If most components have large sizes, only a few nodes are needed as seed URLs (a list of starting pages for a crawler) to be able to reach all the other nodes. If all nodes are members of a single SCC, one URL is enough to crawl all pages.

SCC size	1	2	4	16	19	20	26	45	T[6]
# of SCCs	9	1	3	1	1	1	1	1	18
# of nodes	9	2	12	16	19	20	26	45	149

Table 5. SCC size distribution of the Filipino language community in Indonesia

SCC diameter

The maximum distance between any two nodes is the diameter. For each node size, the diameter is

[5] k = Javanese
[6] Total

calculated and plotted in the chart. Their corresponding SCC graph is also shown.

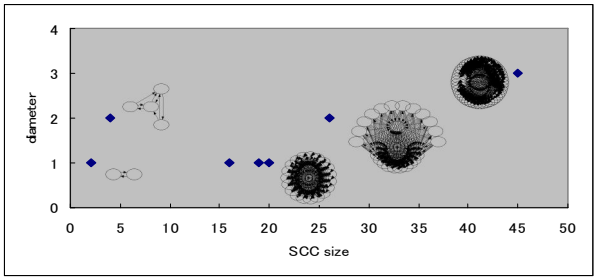

Figure 2. Diameter distribution of the Filipino language community in Indonesia

For the Filipino language subgraph of Indonesia, the component with the largest node size also has the largest diameter. However, the largest diameter size is only 3, which is a very small number. Most of the components have a diameter of 4.

6 Implications

For the Filipino language community, many SCCs can be found in the Philippines, since Filipino is one of its national languages. However, for some countries, there are not many SCCs. For example, Indonesia only has 18 SCCs, half of which consist only of one node. However, the largest component size is 45, and there are 4 more large components. By picking out just one node from each SCC with a large size and using it as a seed URL, many web pages can already be downloaded. Add to it the depth parameter of 3, which is the largest diameter, these web pages can be downloaded within a short period of time.

The choice of seed URL and the crawly depth are useful parameters for crawling. The analysis is done for each language community to get these parameters for language-specific crawling purposes. These parameters are different for each language community. This paper shows only shows the case of the Filipino language community of Indonesia as a sample illustration of the diameter metric.

7 Conclusion

The vastness and multilingual content of the web makes it a rich source of culturally-diversified information. Since web pages are connected by hyperlinks, information can be readily accesses by jumping from one page to another via the hyperlinks.

For each country domain, the web pages written in the same language form a language community. The link structure between language communities shows how connected a language community is with another language community. It can be assumed that the close links between two language communities on the web imply the existence of multilingual speakers of the two languages. Otherwise linked pages will not be visited. In this context, the language graph analysis demonstrated in this study gives an effective tool to understand the linguistic scenes of the country. If the same analysis is performed for the secondary level domain data, further insight into the socio-linguistic status of each language can be drawn. Secondary domain corresponds to different social area of language activities, such as “ac” or “edu” for academic and education arena, “go” or “gov” for government or public arena, and “co” or “com” for commercial business and occupational arena. Although this study does not extend its scope to the secondary level domain analysis, the effectiveness of the approach was demonstrated.

Another implication drawn from this study is that the language graph analysis can identify intermediary languages in the multilingual communities. In the real world, some languages are acting as a medium of communications among the different language speakers. In most cases, such lingua franca are international languages such as English, French, Arabic, etc. But it’s difficult to identify which language is acting as such in detail. But on the web link structure among languages, the language graph can give us a clue to identify this. As shown in this paper, there are a number of languages acting as intermediary between two languages having only a few hyperlinks between them. Although the result of this category is doubtful because of misidentification of language, some cases show the expected result.

The second objective of the study is to give a microscopic level structure of the web communities for much more practical and technical reasons, such as how to design more effective crawling strategy, and how to prepare starting URLs with minimal efforts. The key issue in this context is to reveal the connectedness of the web. To show the connectedness of language communities, several

graph theory metrics, the size and numbers of strongly-connected components and the diameters are calculated and visual presentations of language communities are also given. This information can aid in defining parameters used for crawling, particularly language-specific crawling.

As a summary, the link structure analysis of language graphs can be a useful tool for various spectrums of socio-linguistic and technical research purposes.

8 Limitations and Future Work

The results of this research are highly dependent on the language identification module (LIM). With a more improved LIM, more accurate results can be presented. Currently, there is an ongoing experiment that uses a new LIM.

This analysis will also be done to the secondary-level-domains to show the language distribution for different social areas.

Future work also includes the creation of a language-specific crawler that will incorporate the results derived from the analysis of the SCC size and diameter of the language subgraphs.

Acknowledgment

The study was made possible by the financial support of the Japan Science and Technology Agency (JST) under the RISTEX program and the Asian Language Resource Network Project. We also thank UNESCO for giving official support to the project since its inception.

References

Balakrishnan, Hemant and Narsingh Deo. 2006. Evolution in Web graphs. *Proceedings of the 37th Southeastern International Conference on Combinatorics, Graph Theory, and Computing*. Boca Raton, FL.

Bharat Krishna, Bay-Wei Chang, Monika Henzinger, and Matthias Ruhl. 2001. Who links to whom: mining linkage between Web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51-58, San Jose, California.

Boldi, Paolo, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. 2004. UbiCrawler: A scalable fully distributed web crawler. *Software: Practice & Experience, 34(8):711-726*.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. In *Proceedings of the 9th World Wide Web Conference*, pages 309-320, Amsterdam, Netherlands.

Chakrabarti, Soumen, Martin van den Berg, and ByronDom. 1999. Focused Crawling: a new approach to topic-specific Web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*, pages 1623-1640, Toronto, Canada.

Herring, Susan C., John C. Paolillo, Irene Ramos-Vielba, Inna Kouper, Elijah Wright, Sharon Stoerger, Lois Ann Scheidt, and Benjamin Clark. 2007. Language Networks on LiveJournal. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Hawaii.

Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew S. Tompkins, and Eli Upfal. 2000. The Web as a graph. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 1-10, Dallas, Texas, United States.

Petricek, Vaclav, Tobias Escher, Ingemar J. Cox, and Helen Margetts. 2006. The web structure of e-government: developing a methodology for quantitative evaluation. In *Proceedings of the 15th International World Wide Web Conference*, pages 669-678, Edinburgh, Scotland.

Pingali, Prasad, Jagadeesh Jagarlamudi, and Vasudeva Varma. 2006. WebKhoj: Indian language IR from Multiple Character Encodings. In *Proceedings of the 15th International World Wide Web Conference*, pages 801-809, Edinburgh, Scotland.

Stamatakis, Konstantinos, Vangelis Karkaletsis, Georgios Paliouras, James Horlock, Claire Grover, James R. Curran, and Shipra Dingare. 2003. Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pages 75–78. Edinburgh, UK.

Suzuki, Izumi, Yoshiki Mikami, Ario Ohsato, and Yoshihide Chubachi. 2002. A language and character set determination method based on N-gram statistics. *ACM Transactions on Asian Language Information Processing*, 1(3): 269-278.

Tamura, Takayuki, Kulwadee Somboonviwat, and Masaru Kitsuregawa. 2007. A method for language-specific Web crawling and its evaluation. *Systems and Computers in Japan*, 38(2):10-20.