Fourth Workshop on Algorithms and Models for the Web-Graph (WAW2006)

Nov. 30 – Dec. 1, 2006

Banff International Research Institute (BIRS)
Banff, Alberta

The World Wide Web has become part of our everyday life and information retrieval and data mining on the Web are now of enormous practical interest. The algorithms supporting these activities combine the view of the Web as a text repository and as a graph, induced in various ways by links among pages, links among hosts, or other similar networks.

The aim of the 2006 Workshop on Algorithms and Models for the Web-Graph (WAW2006) is to further the understanding of these Web-induced graphs, and stimulate the development of high-performance algorithms and applications that use the graph structure of the Web. The workshop is meant both to foster an exchange of ideas among the diverse set of researchers already involved in this topic, and to act as an introduction for the larger community to the state of the art in this area.

This is the fourth in a series of very successful workshops on this topic. WAW 2002 and 2004 were held in conjunction with the Annual IEEE Symposium on Foundations of Computer Science (FOCS). WAW2003 was held in conjunction with the Twelfth International World Wide Web Conference.

Proceedings of WAW 2006 will be published by the Springer series Lecture Notes in Computer Science (LNCS).

We gratefully acknowledge the financial support from:

The Mathematics of Information Technology and Complex systems (MITACS)

Network of Centres of Excellence,
Burnaby, British Columbia, Canada
`www.mitacs.math.ca`

Yahoo! Inc.,

Sunnyvale, California,
`www.yahoo.com`

Genieknows.com

Halifax, Nova Scotia
`www.genieknows.com`

Organizing Committee

William Aiello,
Dept. of Computer Science, University of British Columbia, Vancouver, Canada
Andrei Broder,
Yahoo! Inc., Sunnyvale, California
Jeannette Janssen,
Dept. of Mathematics and Statistics, Dalhousie University, Halifax, Canada
Evangelos Milios,
Faculty of Computer Science, Dalhousie University, Halifax, Canada

	Wednesday, November 29
16:00	Check-in begins (Front Desk Professional Development Centre - open 24 hours)
17:30-19:30	Buffet Dinner, Donald Cameron Hall
20:00	Informal gathering in 2nd floor lounge, Corbett Hall
	Opportunity to prepare poster display

	Thursday, November 30
8:00	Opportunity to prepare poster display
8.45	Introduction and Welcome to BIRS by BIRS Station Manager
	Contributed talks — Chair: Pawel Pralat
9:00	Abie Flaxman
	Expansion and lack thereof in randomly perturbed graphs [p. 6]
9:30	Yu Hirate
	Web Structure in 2005 [p. 6]
10:00	Ross Richardson
	Local/global phenomena in geometrically generated graphs [p. 8]
10:30	Anthony Bonato
	Structural properties of infinite limits of self-organizing networks [p. 5]
11:00	Coffee break
11:15	**Keynote speaker**: Fan Chung-Graham
	Local Graph Partitioning using PageRank vectors [p. 3]
12:15	Lunch (Donald Cameron Hall)
14:00	**Keynote speaker**: Soumen Chakrabarti
	Building blocks for semantic search engines: Ranking and compact indexing in entity-relation graphs with associated text [p. 3]
15:00	Coffee break
15:15	**Poster session**
17:30-19:30	Buffet Dinner, Donald Cameron Hall

	Friday, December 1
	Contributed talks — Chair: Anthony Bonato
8:45	Sandro Flammini
	Approximating PageRank from indegree [p. 5]
9:15	Nelly Litvak
	In-degree and Pagerank of Web pages: why do they follow similar power laws? [p. 7]
9:45	Maxim Gurevich
	Search engine sampling via random walks [p. 6]
10:15	Davood Rafiei
	Some applications of snowball sampling on the Web graph [p. 8]
10:45	Coffee break
11:00	**Keynote speaker**: Walter Willinger
	Power laws in Internet graphs: full of sound and fury, signifying nothing? [p. 4]
12:00	Lunch (Donald Cameron Hall)
13:30	**Keynote speaker**: Filippo Menczer
	Googlearchy or Googlocracy? How search affects Web traffic and growth [p. 4]
14:30	Coffee break
14:45	**Contributed talks** — Chair: Nauzer Kalyaniwalla
14:45	Martin Olsen
	Communities in large networks: identification and ranking [p. 7]
15:15	Debora Donato
	Link analysis for Web spam detection [p. 5]
15:45	Igor Kanovsky
	Web graph clustering based on link correlations [p. 7]
16:15	Melih Onus
	A scalable multilevel algorithm for community structure detection [p. 8]
16:45	Closing remarks
17:30-19:30	Buffet Dinner, Donald Cameron Hall

1 Abstracts / bios of keynote speakers

Local Graph Partitioning using PageRank Vectors

Reid Andersen, ***Fan Chung*** *and Kevin Lang*
UC San Diego

A local graph partitioning algorithm finds a cut near a specified starting vertex, with a running time that depends largely on the size of the small side of the cut, rather than the size of the input graph. We give a local partitioning algorithm using a variation of PageRank with a specified starting distribution. We derive a mixing result for PageRank vectors similar to that for random walks, and show that the ordering of the vertices produced by a PageRank vector reveals a cut with small conductance. In particular, we show that for any set C with conductance Φ and volume k, a PageRank vector with a certain starting distribution can be used to produce a set with conductance $O(\sqrt{\Phi \log k})$. We present an improved algorithm for computing approximate PageRank vectors, which allows us to find such a set in time proportional to its size. By combining small sets found by this local partitioning algorithm, we obtain a cut with conductance ϕ and approximately optimal balance in time $O(m \log^4 m/\phi^2)$.

Fan Chung Graham received a B.S. degree in mathematics from National Taiwan University in 1970 and a Ph.D. in mathematics from the University of Pennsylvania in 1974, after which she joined the technical staff of AT&T Bell Laboratories. From 1983 to 1991, she headed the Mathematics, Information Sciences and Operations Research Division at Bellcore. In 1991 she became a Bellcore Fellow. In 1993, she was the Class of 1965 Professor of Mathematics at the the University of Pennsylvania. Since 1998, she has been a Professor of Mathematics and Professor of Computer Science and Enginering at the University of California, San Diego. She is also the Akamai Professor in Internet Mathematics.
Her research interests are primarily in graph theory, combinatorics, and algorithmic design, in particular in spectral graph theory, extremal graphs, graph labeling, graph decompositions, random graphs, graph algorithms, parallel structures and various applications of graph theory in Internet computing, communication networks, software reliability, chemistry, engineering, and various areas of mathematics. She was awarded the Allendoerfer Award by Mathematical Association of America in 1990. Since 1998, she has been a member of the American Academy of Arts and Sciences.

Building blocks for semantic search engines: Ranking and compact indexing in entity-relation graphs

Soumen Chakrabarti
IITB, Mumbai

We see an evolutionary path to supporting semantic search over text facilitated by 1. extractors and annotators for ever-growing collections of entity and relation types and 2. search systems that exploit a smooth continuum between structured entities and relations on one hand and uninterpreted text on the other. The extractors and annotators will be imperfect and incomplete. Unlike in traditional data warehousing, the unstructured source cannot be forgotten once structured data is curated. The source text lives on, with annotations as probabilistic connections to one or more ontologies. Queries involve ontology elements as well as uninterpreted strings. Searchers know only bits and pieces of schema. The query language must enable schema-free searches but reward schema knowledge. In such a system, ranking of results cannot be made all explicit as in SQL. In the first part of the talk I visit two issues related to scoring and ranking results. In one scenario we wish to rank mentions of instances of a given type (e.g. distance) based on their textual proximity to query keywords (e.g. Paris Helsinki). Such queries are crude but effective filters for simple questions like "what is the distance between Paris and Helsinki". In the other scenario I generalize the setting to learning ranking functions in arbitrary E-R graphs, given partial order preferences. In the second part of the talk I discuss the difficulties faced in indexing and query processing. I describe new techniques to estimate query-processing cost and cost-driven index compaction based on query logs. The work described is embodied in a system we are building that we plan to release in the public domain.

Soumen Chakrabarti received his B.Tech in Computer Science from the Indian Institute of Technology, Kharagpur, in 1991 and his M.S. and Ph.D. in Computer Science from the University of California, Berkeley in 1992 and 1996. At Berkeley he worked on compilers and runtime systems for running scalable parallel scientific software on message passing multiprocessors.
He was a Research Staff Member at IBM Almaden Research Center from 1996 to 1999, where he worked on the Clever Web search project and led the Focused Crawling project.
In 1999 he joined the Department of Computer Science and Engineering at the Indian Institute of Technology, Bombay, where he has been an Associate professor since 2003. In Spring 2004 he was Visiting Associate professor at Carnegie-Mellon University.

He has published in the WWW, SIGIR, SIGKDD, SIGMOD, VLDB, ICDE, SODA, STOC, SPAA and other conferences as well as Scientific American, IEEE Computer, VLDB and other journals. He holds eight US patents on Web-related inventions. He has served as technical advisor to search companies and vice-chair or program committee member for WWW, SIGIR, SIGKDD, VLDB, ICDE, SODA and other conferences, and guest editor or editorial board member for DMKD and TKDE journals. He is also author of a book on Web Mining.
His current research interests include integrating, searching, and mining text and graph data models, exploiting types and relations in search, and Web graph and popularity analysis.

Power laws in Internet graphs: Full of sound and fury, signifying nothing?

Walter Willinger
AT&T Labs-Research
walter@research.att.com

Internet measurements have been a rich source for discovering power law relationships. However, many of these power law "discoveries" have turned out to be specious, especially with respect to connectivity-related measurements that are notorious for their ambiguities, incompleteness, and inaccuracies. This is bad news for many of the modeling efforts that have focused almost exclusively on reproducing the claimed power law relationships. In this talk, I will discuss an alternate modeling approach that is not tied to any claimed power law behavior and instead relies heavily on domain knowledge. This approach is capable of explaining a wide range of different system behaviors and provides a basis for exploring under which circumstances one should or should not expect power law relationships in Internet connectivity structures such as the Internet's router-level topology, AS graph, or overlay networks such as the Web or different Peer-to-Peer networks.

Walter Willinger received the Diplom (Dipl. Math.) from the ETH Zurich, Switzerland, and the M.S. and Ph.D. degrees from the School of ORIE, Cornell University, Ithaca, NY. He is currently a member of the Information and Software Systems Research Center at AT&T Labs - Research, Florham Park, NJ, and before that, he was a Member of Technical Staff at Bellcore Applied Research (1986-1996). His research interests include studying the multiscale nature of Internet traffic and topology and developing a theoretical foundation for dealing with large-scale communication networks such as the Internet. He is a Fellow of ACM (2005) and a Fellow of IEEE (2005). For his work on the self-similar ("fractal") nature of Internet traffic, he received the 1996 IEEE W.R.G. Baker Prize Award, the 1994 W.R. Bennett Prize Paper Award, and the 2005 ACM/SIGCOMM "Test of Time" Paper Award.

Googlearchy or Googlocracy? How search affects Web traffic and growth

Filippo Menczer
Indiana University

Search engines have become key media for our scientific, economic, and social activities by enabling people to access information on the Web in spite of its size and complexity. On the down side, search engines bias the traffic of users according to their page-ranking strategies, and some have argued that they create a vicious cycle that amplifies the dominance of established and already popular sites. We show that, contrary to these prior claims and our own intuition, the use of search engines actually has an egalitarian effect. We reconcile theoretical arguments with empirical evidence showing that the combination of retrieval by search engines and search behavior by users mitigates the attraction of popular pages, directing more traffic toward less popular sites, even in comparison to what would be expected from users randomly surfing the Web. We then extend the analysis of traffic to a general model of search-driven network growth, that predicts the topological propertied of the Web graph. Joint work with Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani.

Filippo Menczer is an associate professor of informatics and computer science, adjunct associate professor of physics, and a member of the cognitive science program at Indiana University, Bloomington. He holds a Laurea in Physics from the University of Rome and a Ph.D. in Computer Science and Cognitive Science from the University of California, San Diego. Dr. Menczer has been the recipient of Fulbright, Rotary Foundation, and NATO fellowships, and is a fellow-at-large of the Santa Fe Institute. His research is supported by a Career Award from the National Science Foundationon and focuses on Web, text, and data mining, Web intelligence, distributed information systems, social Web search, adaptive intelligent agents, complex systems and networks, and artificial life.

List of abstracts of presentations

(Ordered alphabetically by speaker's last name)

Structural properties of infinite limits of self-organizing networks

Anthony Bonato *and Jeannette Janssen*
Department of Mathematics, Wilfrid Laurier University, Waterloo ON, Canada
Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada
abonato@rogers.com, janssen@mathstat.dal.ca

One of the most widely studied real-world networks is the web graph, whose nodes represent web pages, and whose edges represent the links between pages. Another well-studied real-world network consists of protein-protein interactions in a living cell. Both networks are *self-organizing*, as each node acts as an independent agent that bases its decision on how to link to the existing network on local knowledge. Many models for self-organizing networks are *on-line*: new nodes are born over time. Hence, it is natural to consider the infinite graphs that result in the limit as time tends to infinity.

We present new results on limit graphs generated by on-line random graph models. We will characterize properties of a generalized copying model via adjacency properties, and describe self-similarity properties of limit graphs generated by on-line processes. If time permits, then we will discuss a new geometric model for the web graph.

Link Analysis for Web Spam Detection

Luca Becchetti, Carlos Castillo, ***Debora Donato****, Stefano Leonardi and Ricardo Baeza-Yates*
Universitá di Roma "La Sapienza", Roma, Italy
Yahoo! Research, Barcelona, Spain & Santiago, Chile

Since its inception, the Web has been considered as a privileged means for free exchange of information and resources among users. Unfortunately this feature has been revealed also its major drawback, since it has catalyzed the (more or less legitimate) interests of ventures seeking for easy profits against low investments and risks.

Since most of the informational and transactional needs of people all over the world are satisfied by querying search engines, there is an economic incentive for manipulating search engines indexes in order to allow pages to get an undeserved high score, a phenomenon nowadays well known as **search engine spamming** or **spamdexing**.

In this work we presented an algorithmic approach for spam detection based on link analysis. This methodology results effective against **topological spamming**, a form of spamming that acts creating densely connected set of pages, called **spam farms**, able to modify the local properties of the pages. We study several metrics not considered before to build an automatic classifier: we test in our collection TrustRank, an algorithm able to estimate the amount of score that a page receive from trust sources. Moreover we propose the use of degree-degree correlations, edge-reciprocity and above all we adapt rank propagation and probabilistic counting algorithms to identify spam farms.

Approximating PageRank from indegree

S. Fortunato, M. Boguna, ***A. Flammini****, F. Menczer*
School of Informatics Indiana U., Uni. Bielefeld, Uni. Barcelona

PageRank is a key element in the success of search engines, allowing to rank the most important hits in the top screen of results. One key aspect that distinguishes PageRank from other prestige measures such as in-degree is its global nature. From the information provider perspective, this makes it difficult or impossible to predict how their pages will be ranked. Consequently a market has emerged for the optimization of search engine results. Here we study the accuracy with which PageRank can be approximated by in-degree, a local measure made freely available by search engines. Theoretical and empirical analyses show that given the weak degree correlations in theWeb link graph, the approximation can be relatively accurate, giving service and information providers a new marketing tool.

Expansion and lack thereof in randomly perturbed graphs

Abraham D. Flaxman
Microsoft Research, USA

This extended abstract examines the expansion properties of randomly perturbed graphs. These graphs are formed by, for example, adding a random 1-out or very sparse Erd?os-Renyi graph to an arbitrary connected graph. It is shown that there exists a constant such that when any connected n-vertex base graph G is perturbed by adding a random 1-out then, with high probability, the resulting graph has e(S, S) —S— for all S V with —S— 3 4n. The analogous statement for perturbations by Gn, /n is also considered, and under this perturbation, the expansion of the perturbed graph depends on the structure of the base graph. A necessary and sufficient condition for the base graph is given under which the resulting graph is an expander with high probability. These techniques are also applied to study expansion and rapid mixing in the small worlds graphs described by Watts and Strogatz in [Nature 292 (1998), 440-442] and by Kleinberg in [Proc. of 32nd ACM Symposium on Theory of Computing (2000), 163-170]. Analysis of Kleinberg's model shows that the graph stops being an expander exactly at the point where a decentralized algorithm is effective in constructing a short path.

Search Engine Sampling via Random Walks

Ziv Bar-Yossef and ***Maxim Gurevich***
Department of Electrical Engineering, Technion, Haifa 32000, Israel.
Google Haifa Engineering Office, Haifa, Israel.
gmax@tx.technion.ac.il, zivby@ee.technion.ac.il

In our WWW2006 paper we proposed two algorithms for sampling documents from a corpus of documents indexed by a search engine, using only the engine's public interface. One of the algorithms is based on a random walk on the corpus. The algorithm employs an "approximate" variant of the Metropolis-Hastings (MH) algorithm to transform a simple random walk on the corpus into a Markov chain that converges to any desired target distribution over the corpus. We studied the quality and the performance of the algorithm only empirically.

Here we provide rigorous theoretical analysis of the bias and the efficiency of the algorithm and show that while its bias is reasonable, the algorithm requires a prohibitively large number of queries to generate each sample. We then propose a generalization of the Maximum Degree (MD) method as an alternative to the Metropolis-Hastings algorithm. We analyze the new algorithm and prove that it has exactly the same bias as the MH-based algorithm, while it can be significantly more efficient.

The generalized MD method and the analysis we provide for the approximate variants of the MH algorithm and the MD method may be of independent interest.

Web Structure in 2005

Yu Hirate*, Kato Shin and Hayato Yamana*
Dept. of Computer Science, Waseda University, Japan
Mitsubishi Electric Corporation, Japan
{hirate,kato,yamana}@yama.info.waseda.ac.jp

The number of static web pages is estimated over 15 billion in Oct 2005. This is multiplying 200 pages by 74.4 million web servers, where 200 pages means the average number of web pages and are assumed from past three researches. However, based on the analysis of 8.5 billion web pages that we have crawled by Oct. 2005, we estimate the total number of web pages as 40 billion. This is because dynamic web pages have increased rapidly in recent years.

We also constructed the web structure based on 3 billion web pages in 2005. Then compared the web structure constructed by Broder *et al.* in 1999 to our web structure. As a result, we figure out that the size of "CORE", the center component of bow-tie structure, is increasing in recent years, especially in Chinese and Japanese web.

Web Graph Clustering Based on Link Correlations

Igor Kanovsky
Max Stern Academic College of Emek Yezreel, Israel
igork@yvc.ac.il

In this paper, we describe a new method for cluster recognition in complex networks and its application to the Web graph. A typical network has Small World properties or/and some other edges correlation. The method invokes link weighting based on the link participants in local links correlation. Inter-cluster links are perceived as not correlated or weak correlated. We propose an extension of the popular Small World model of Watts and Strogatz and use it to test our approach.
The proposed approach has a set of advantages. Efficiency: has a polynomial complexity. Locality: no need to know all the data of the graph or number of clusters for local cluster recognition. Applicability: applicable for graphs of a different nature (”big” and ”small”, power-law, etc.). It helps to recognize Small World sub-graphs within a non-Small World graph. The method may be extended for different networks by adopting the weighting mechanism.

In-Degree and PageRank of Web pages: Why do they follow similar Power Laws?

N. Litvak, *W.R.W. Scheinhardt and Y. Volkovich*
Dept. of Applied Math, University of Twente, Netherlands

The PageRank is a popularity measure designed by Google to rank Web pages. Experiments confirm that the PageRank obeys a ‘power law’ with the same exponent as the In-Degree. This paper presents a novel mathematical model that explains this phenomenon. The relation between the PageRank and In-Degree is modelled through a stochastic equation, which is inspired by the original definition of the PageRank, and is analogous to the well-known distributional identity for the busy period in the $M/G/1$ queue. Further, we employ the theory of regular variation and Tauberian theorems to analytically prove that the tail behavior of the PageRank and the In-Degree differ only by a multiplicative factor, for which we derive a closed-form expression. Our analytical results are in good agreement with experimental data.

Communities in Large Networks: Identification and Ranking

Martin Olsen
Department of Computer Science, University of Aarhus
mo@daimi.au.dk

We study the problem of identifying and ranking the members of a community in a very large network with link analysis only, given a set of a (few) representatives of the community.
We define the concept of a *community* justified by a formal analysis of a simple model of the evolution of a directed graph. We show that the problem of deciding whether non trivial communities exists is NP complete. Nevertheless experiments show that a very simple greedy approach can identify members of a community in the Danish part of the www graph with time complexity only dependent on the size of the found community and its immediate surroundings.
We rank the members in a community by performing a computationally inexpensive calculation which is a “local” variant of the PageRank algorithm. The mathematical model behind the ranking is a small Markov Chain with the community as its state space forming a valuable basis for analyzing consequences of changes of the link structure.
Results are reported from successful experiments on identifying and ranking Danish Computer Science sites and Danish Chess pages using only a few representatives.

A scalable multilevel algorithm for community structure detection

Hristo N. Djidjev and ***Melih Onus***
Los Alamos National Laboratory, Los Alamos, NM
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ

One way to analyze and understand the information contained in the huge amount of data available on the WWW and the relationships between the individual items is to organize them into ”communities,” maximal groups of related items. Determining the communities is of great theoretical and practical interest since they correspond to entities such as collaboration networks, online social networks, scientific publications or news stories on a given topic, related commercial items, etc. Communities also arise in other types of networks such as computer and communication networks (the Internet, ad-hoc networks) and biological networks (protein interaction networks, genetic networks).

The problem of identifying communities in a network is usually modeled as a *graph clustering* problem, where vertices correspond to individual items and edges describe relationships. Then the communities correspond to subgraphs with dense connections between vertices from the same subgraph and fewer connections between vertices in different subgraphs. The graph clustering problem has been intensively studied in the recent years, but the algorithms reported in the literature are typically either not fast enough, or inaccurate.

In this study we will describe a new algorithm for community detection that uses a newly discovered relationship between the graph clustering and the graph partitioning problems. Specifically, we will show that an important measure of the quality of graph clustering, called modularity, can be optimized by solving a graph partitioning problem. The partitioning problem itself is solved using a multilevel procedure, resulting in a community detection algorithm that is both accurate and scalable.

Some applications of snowball sampling on the Web graph

Davood Rafiei
Department of Computing Science, Univ. of Alberta

Searching the Web graph has much in common with collecting hard-to-reach subjects in social networks. Both perform link tracing to reach subjects being studied or pages that satisfy a query. In this talk, we present the links between snowball sampling and some of the search algorithms on the Web. We also discuss our work on sampling from the Web graph and visualizing its connectivity structure.

Local/Global Phenomena in Geometric Random Graphs

Ross M. Richardson
University of California, San Diego

Many real-world graphs have a natural notion of connections being either local or global. Phone networks, for instance, consist of both local grid lines as well as long trunk lines. To study the distribution of both local and global connections in complex networks, we consider a geometric random tree model in which each edge arises from an optimization rule reflecting a distance-to-center cost. In particular, we show how varying the optimization rule causes the tree to shift from a global, star-like configuration to a purely local configuration. There is a sharp transition between these two behaviors, and in between we find a more complex random tree where neither behavior dominates. Finally, we examine the stability of this behavior and how it interacts with varying notions of locality.

List of abstracts of posters

(Ordered alphabetically by presenter's last name)

Combating Spamdexing: Incorporating Heuristics in Link-based Ranking

Tony Abou-Assaleh *and Tapajyoti Das*
GenieKnows.com, Halifax, Nova Scotia, Canada
{taa,tdas}@genieknows.com

Spamdexing is an activity that attempts to artificially manipulate a page's ranking in a search engine. Spamdexing renders search engines ineffective because it targets and compromises one of the main benefits of search engines: retrieving relevant results.

Our algorithm starts with cleaning up the link graph by removing all internal and bidirectional links. Next, we use simple heuristics to identify a core set of Web spam pages. This set is further extended to include other likely spam pages using a Spam Propagation algorithm. Finally, we use a biased PageRank-based ranking algorithm to produce the final off-line scores.

We manually evaluate the effectiveness of this approach in combating spamdexing. We observe only 9% of the top 100 pages are spam in the biased PageRank, giving a dramatic reduction of 33% over the baseline, in which 42% of the top 100 pages are spam.

Neighborhood-conscious Clustering

Ralitsa Angelova*, Stefan Siersdorfer, Gerhard Weikum*
Max-Planck Institute for Informatics, Saarbrücken, Germany
{angelova, stesi, weikum}@mpi-inf.mpg.de

We present a flexible framework which addresses the problem of automatically clustering linked document collections. The proposed algorithm starts with an initial clustering, emphasizes the strongest inter-objects relationships built into a first-order Markov Random Field model, and uses an iterative relaxation labeling technique to redistribute objects among clusters based on the cluster label assignments in the objects' neighborhoods. The framework can be built on top of any clustering algorithm to produce a more robust and higher-quality clustering result.

Towards Adaptive Web Search Engines

M. Barouni-Ebrahimi *and Ali A. Ghorbani*
Faculty of Computer Science, University of New Brunswick, Canada,
{m.barouni;ghorbani}@unb.ca

Web search engines efficiently surf the Internet and return the most relevant pages to the users' queries. However, the order of the recommended pages is not always in accordance with the users' priorities. The users needs to check the list of the recommended pages to find one of their interests. On the other hand, the queries sent by the users do not always corresponds to their intentions. The lack of user knowledge or unfamiliarity with the specific keywords and phrases in the domain knowledge leaves the user wondering about what phrases would be the most related ones to his desire. The contribution of our research is three folds. First, Complementary Phrase Recommender module suggests to the user a list of complementary phrases for his uncompleted query. Second, Related Phrase Advisor module provides a list of phrases related to the query segment that user has entered. These two modules guide the user to enter the more related phrases to his intention as a query. Third, Page Rank Revisor module refines the order of the recommended documents prepared by a conventional web search engine to help the user find the related web pages at top of the list.

Categorization of graphs using k-cores

John Healy*, J. Janssen, E. Milios, W. Aiello*
Dalhousie University, University of British Columbia

A k-core of a graph is the subgraph generated by recursively removing all nodes with degree $< k$. This can be thought of as a weaker version of a clique. k-cores are useful for pruning low importance vertices. ncreasing the value k eliminates nodes resulting in a component either remaining in our graph, splitting into multiple components, or being eliminated entirely. The resulting small directed acyclic graph, capturing the evolution of components as k increases, reveals the structure of a graph. We are using the k-core representation in order to find which of several available generative models is the best description of a real world graph, by developing a method of summarizing the component trees for statistical comparison.

Random k-Tree as a Model for Complex Networks

Y. Gao and ***C. Hobson***
Univ. of British Columbia Okanagan

Since the discovery of the power-law degree distribution of web graphs and other complex networks, many random models with a power-law distribution have been proposed and intensively studied. Most of these models are based on the preferential attachment mechanism to generate graphs with a power-law degree distribution. In this note, we discuss our ongoing work on understanding a natural random distribution of a well-known class of graphs, namely the k-trees and partial k-trees, which may also serve as an alternative and viable model for complex networks. The notion of a k-tree can be regarded as a generalization to that of a tree and is closely related to the concept of treewidth in graph theory. As many NP-hard problems can be solved polynomially on graphs with a fixed treewidth (i.e., partial k-trees with k fixed), we believe this class of k-tree based random models has the potential to better capture some algorithmically-relevant features that have been largely ignored in existing models for complex networks. Random (partial) k-tree also poses some theoretical questions that are of interest in their own right.

How NAGA uncoils: Searching with Relations and Entities

Gjergji Kasneci*, Maya Ramanath, Fabian Suchaneck and Gerhard Weikum*
Max-Planck-Institut für Informatik, Saarbrücken / Germany

The everlasting enrichment of the Web with certain as well as uncertain and unstructured information calls for vertical search techniques which fulfill users' needs for querying the Web in a more precise way. Going one step beyond keyword search and allowing the specification of contextual concepts for keywords or relationships holding between them clears the way for new attractive and promising possibilities.

We present NAGA: A semantic search engine for the Web which exploits relationships between entities for precise query specification and answering. NAGA's trump card is its ontological knowledge graph built on top of a refined data model which in turn serves as a basis for NAGA's query model and answer computation algorithms. NAGA extracts facts from Web pages and stores them into the above mentioned knowledge graph. Not only the extracted facts are recorded, but also a confidence measure for each fact is computed and maintained. NAGA provides a query language which can be capable of expressing queries ranging from simple keyword queries to complex graph queries which utilize regular expressions over relation names. NAGA's answer model is based on subgraph matching algorithms which in turn make use of intuitive scoring and ranking mechanisms. The approach we follow represents a general approach towards the semantic processing of information extracted from any unstructured text corpora.

An Approach to Web-Site Summarization by Image Content

E. Baratis, E. Petrakis and ***E. Milios***
Dept. of Comp. Eng. Tech. Univ. of Crete, Fac. Of Comp. Sci. Dalhousie U.

Image-based abstraction (summarization) of a Web site relates to extracting the most characteristic (or important) images from it. This process is complementary to text summarization which works by extracting brief text summaries (e.g., important phrases or sentences) from the Web sites. The proposed method works by analyzing the content of the images stored in a Web-site taking also link information into account. Because there is no unique method for analyzing the content of every possible image type, this work focuses on logo and trademark images. The method incorporates machine learning for distinguishing logo and trademarks from images of other categories (e.g., landscapes, faces). Because the same logo or trademark may appear many times in various forms within the same Web site, only unique logo and trademark images are extracted. These images are then ranked by importance. The most important logos and trademarks are finally selected to form the image summary of a Web site.

Growing and classical protean graphs (new probabilistic models of the web)

Pawel Pralat
Department of Mathematics and Statistics, Dalhousie, University, Halifax NS, Canada B3H 3J5
pralat@mathstat.dal.ca

The web may be viewed as a graph each of whose vertices corresponds to a static HTML web page, and each of whose edges corresponds to a hyperlink from one web page to another. Recently there has been considerable interest in using random graphs to model complex real-world networks to gain an insight into their properties.

We propose an extended version of a new random model of the web graph in which the degree of a vertex depends on its age. We use the differential equation method to obtain basic results on the probability of edges being present. From this we are able to characterize the degree sequence of the model and study its behaviour near the connectivity threshold.

We present also the classical version of the model and characterize the limit distribution of the 'recovery time' for connectivity near the connectivity threshold, and the diameter of the giant component.

This is a joint work with Tomasz Luczak and Nicholas Wormald.

Traps and Pitfalls of Order-Based Correlation Indices for Topic-Biased PageRank

*Paolo Boldi, **Roberto Posenato**, Massimo Santini and Sebastiano Vigna*
Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano, Milano, Italy.
Dipartimento di Informatica, Università degli Studi di Verona, Verona, Italy

Several studies have been recently published about the correlation between rank vectors for the same web graph obtained with different ranking techniques, or computed with bias towards different topics.

The first contribution of our work is a publicly available snapshot of the `.uk` domain, together with topic bias data derived from the ODP hierarchy. We believe such a public, well-defined data set is essential to continue research on topic-biased ranking.

After, we extend the closed formula given by Del Corso, Gullì and Romani for strongly preferential PageRank to a general formula that applies also to weakly preferential PageRank. Using this formula, any biased, weakly preferential PageRank vector whose distributions are a linear combination of a set of base vectors can be computed using the *pseudorank* vectors associated to the base vectors.

The last issue we tackle, however, is probably the most interesting one. Correlation measures such as Kendall's τ are based on the number of exchanges appearing in the rank list. The point that appears to have been completely missed in the literature is that the computation of ranks is almost always the result of interrupting a limiting process (e.g., the power method).

As a result, a number of correct digits appearing in the ranks is unpredictable, as it just depends on the computational process. The common norms used to interrupt the process guarantee on *average* a certain number of significant digits.

In the case several very close values appear in the ranking list, the effect of such an unpredictable precision turns out to be catastrophic.

We have experimentally proved that the τ_b of a certain rank vector *computed against the same vector, but with a different precision* can go down as low as 0.2. Of course, as far as the computation of τ_b uses no more digits than those that are guaranteed to be correct, the correlation is 1, but it rapidly drops as soon as more digits are considered; in particular, computing τ_b blindly can bring essentially to random results.

Web Structure Mining by Isolated Stars

***Yushi Uno**, Yoshinobu Ota, Akio Uemichi and Motohide Umano*
Department of Mathematics and Information Sciences,
Graduate School of Science, Osaka Prefecture University, Japan
uno@mi.s.osakafu-u.ac.jp

In the explosively evolving Web, by regarding the Web as a huge database, it is extremely important not only to obtain primary information but to find hidden information that cannot be found by naive retrievals. This is often called 'web mining', and web structure mining usually aims to find hidden communities that share common interests in specified topics in the Web by focusing on the webgraph that represents the link structure among web pages. In this paper, we newly identify a typical frequent substructure by observing the webgraph, and define it as an isolated star (i-star) so that it becomes easy to be enumerated. We then propose an efficient enumeration algorithm of i-stars, and try structure mining by enumerating them from the real web data. As a result, we observed that most of i-stars correspond to index structures in single domains, while some of them are verified to stand for useful communities. This implies the validity of i-stars as candidate substructure for structure mining.

Representing and Quantifying Rank-change for the Web Graph

Akrivi Vlachou, *Michalis Vazirgiannis and Klaus Berberich*
Department of Informatics, Univ. of Economics and Business, Athens, Greece
Gemo, INRIA, Paris France
Max-Planck-Institut für Informatik, Saarbrücken / Germany
{avlachou, mvazirg}@aueb.gr, kberberi@mpi-sb.mpg.de

The web is a dynamic structure that is constantly changing. The evolution of the web graph is mainly caused by the changes in the graph structure and in the content of the web pages. Every day the web increases both in terms of new pages and new links that interconnect them. One of the biggest challenges is that of searching these vast amounts of data. The research area of web search inherently involves the issue of page ranking. Thus, we claim that the changes in the graph structure is of higher importance as those predominantly cause the changes in authority score and therefore of the web page ranking. In this paper we address the issue of representing and quantifying the web graph evolution. Since a dominant issue is the ranking of pages we define the *rank change rate (racer)* quantifying the web graph evolution.
Our approach of *web graph representation* through the *racer* values enables a concise representation of the web graph - in terms of keeping only the changes among snapshots. Therefore, it is possible to answer as-of queries for the past, to identify trends and so to predict future trends in the web graph. Using our proposed schema for quantifying the web graph evolution the web graph can be represented either as a Markov model and making *predictions* for future ranking values or via piecewise linear approximations towards "as of" queries for the past.
To summarize, in this paper we address the issue of representing and quantifying the web graph evolution. The key contributions of this paper are:

- We propose a rank normalization method and present *racer*, a metric that calculates the ranking change rate of the web pages. We generate *normalized racer* sequences which are used in order to describe the evolution of the web.
- We discuss the problem of finding the dynamic parts of the graph as those that change fast. In addition, we pose the problem of finding aggregate trends in the graph, i.e. how much does a graph related to a query term changes. Moreover, we outline the possible correlation between graphs and trends.
- We present some - due to space limitations - initial experiments of the proposed metrics and estimate the expressiveness of our proposed method to describe the evolution of the web graph.

Sketching Landscapes of Page Farms

Bin Zhou *and J. Pei*
Simon Fraser University, Canada
bzhou@cs.sfu.ca

Ranking pages is an essential task in web search. For a web page p, what other pages are the major contributors to the ranking score of p? How are the contributions made? Understanding the general relations of web pages and their environments is important with a few interesting applications such as spamming detection and community identification and analysis.
In this paper, we study a novel web mining problem: mining page farms and its application in link spamming detection. A page farm is the set of pages contributing to (a major portion of) the PageRank score of a target page. We show that extracting page farms is computationally expensive, and propose heuristic methods. We investigate how to use page farms in link spamming detection. Using a real sample of more than 3 million web pages, we analyze the statistics of "landscapes" of page farms. We examine the effectiveness of our method using a newly available real data set of spamming pages. The empirical study results strongly indicate that our method is effective.