Advances in Link Analysis

Introduction to Information Retrieval CS 221
Donald J. Patterson



Advances in Link Analysis

"The Impact of Crawl Policy on Web Search Effectiveness" by Fetterly, Craswell, Vinay SIGIR2009

"Link Analysis for Private Weighted Graphs" by Sakuma, Kobayashi SIGIR2009

The Impact of Crawl Policy on Web Search Effectiveness

Dennis Fetterly Microsoft Research Mountain View, CA USA fetterly@microsoft.com

Nick Craswell Microsoft Research Cambridge, UK nickcr@microsoft.com

Cambridge, UK vvinay@microsoft.com

Crawl selection policy has a direct influence on Web search effectiveness, because a useful page that is not selected for crawing will also be absent from course results. Yet shows here there are the course to be absent from course process. also be absent from search results. Yet there has been little or no work on measuring this effect. We introduce an evaluation framework, based on relevance judgments pooled from multiple search engines, measuring the maximum potential NDCG that is achievable engines, measuring the maximum potential NDCG that is achievable using a particular crawl. This allows us to evaluate different crawl policies and investigate important scenarios like selection stability poncies and investigate important scenarios like selection stability over multiple iterations. We conduct two sets of crawling experiover multiple iterations. We conduct two sets of crawing experi-ments at the scale of 1 billion and 100 million pages respectively. ments at the scare of 1 minor and 100 minor pages respectively.

These show that crawl selection based on PageRank, indegree and trans-domain indegree all allow better retrieval effectiveness than a simple breadth-first crawl of the same size. PageRank is the most reliable and effective method. Trans-domain indegree can outperform PageRank, but over multiple crawl iterations it is less effective and regenerity, our ever manager craws measures it is less effective and more unstable. Finally we experiment with combinations of crawl and the combination of crawless and the combination of crawless and the combination of crawless and combination of crawle more unstable. Finally we experiment with combinations of crawl selection methods and per-domain page limits, which yield crawls with greater potential NDCG than PageRank.

Categories and Subject Descriptors H.3.3 [Information Storage and Retrieval]: Information Search

and Retrieval

General Terms

Measurement, Experimentation

A useful Web search result will only be seen by users if it is 1. INTRODUCTION erawled by the search engine, indexed correctly, found in the index crawieu by the search engine, indexed correctly, iound in the index when matched with a query and ranked highly in the search result. listing. It only takes one failure in this chain of events for the useful issing. It only taxes one tanire in his crain of events for the userial (relevant) result to be lost. If such failures happen often, users

will perceive a drop in the quality of search results. Therefore, to will perceive a crop in the quarry of search results. Therefore, to optimize user satisfaction, it is important to avoid failure at every stage.

Success at the crawling stage depends on the size of the crawl and the crawl selection policy. For example, the policy of preferring

Permission to make digital or hard copies of all or part of this work for personnal or electronic tree is presented without fee perceided that coming are Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to bear this notice and the full citation or to redistribute to lists, requires prior specific

pages with highest PageRank [7] and a size limit of N leads to pages with nighest PageRank 1/1 and a size firmt of N leads to selecting a set of N high-PageRank pages. When searches are carried out, the quality of search results will sometimes be reduced carried out, the quality of search results will sometimes be reduced because pages that would have been relevant and retrievable were not selected for crawling. One way to reduce such failures is to not selected for crawling. One way to reduce such failures is to increase the size N of the crawl. Another approach is to improve

Although well-known methods exist for evaluating search rel-Annough well-known memous exist for evaruating search reference, such as NDCG [13], we are not aware of any published evance, such as NLCAU [13], we are not aware or any punnsned experiments that compare the relevance achievable by different experiments that compare the relevance acmevante by different crawl policies. Acting as a barrier to experimentation are the large communication and computational costs of conducting multiple communication and computational costs of considering natural crawls, creating multiple indices and processing queries. Our framework appealment to the plant appealment and appealment and appealment appealment appealment and appealment appealm work ameliorates this via a crawl sandbox and an evaluation metric work amenorates this via a crawi santonox and an evaluation metre that only requires the set of selected URLs. The sandbox is simply and the sandbox is said to the said to the sandbox is said to the sa a cache, to avoid crawling URLs more than once if selected by multiple policies or iterations. The metric, maxNDCG, is the best numple policies or nerations. The meme, many code, is the seed potential NDCG that could be achieved based on the presence or absence of relevant pages in a crawl, maxNDCG is proportional to NDCG but may be calculated without indexing and retrieval. It may NIA-U out may or carcurated without indexing and redieval. It may even be calculated for a selected set of pages without attempting to even or carcurated for a selected set of pages without attempting a crawl them, estimating the NDCG that would be achievable by perfect ranker if all selected pages were successfully crawled. These efficiency techniques allow us to run a large numb

experiments comparing crawl policies. We focus on policy experiments comparing craws principes. We tocal on principles selecting a new crawl based on the link graph of a previous cr Seaccung a new craws based on one and grape of a previous ca This is a common scenario, allowing an engine to shift its s towards pages that are preferred according to some link-bay (such as PageRank) but not yet included in the crawl. In Section 2, we discuss different aspects involved in

in occuon 2, we ensure uniferent aspects involved in tion of crawling methods. We provide motivation for or tal setup, and where appropriate, we provide referen work. We then present experiments in Section 3 and

2. CRAWLING AND EVALUAT Search engines are the primary discovery to on the Web, and the Web has an effectively on use wee, and the wee has an effectively that might be indexed. A search engine mus pages it indexes, to make the best use of its Search engines use crawlers to download which indices are built. Starting from a download pages and extract links. It rec seen and adds any unseen URLs to a s After an initial crawl, there is the

corpus, since pages are continually deleted [4]. One option would be to although completely throwing ay

Link Analysis for Private Weighted Graphs

University of Tsukuba 1-1-1 Tennodai, jun@cs.tsukuba.ac.jp

ABSTRACT

Link analysis methods have been used successfully for knowledge Little analysis memous nave been used successionly for knowledge discovery from the link structure of mutually linking entities. Exdiscovery from the link structure of mutually linking entities. Ex-isting link analysis methods have been inherently designed based on the fact that the entire link structure of the target graph is obon the fact that the entire and structure of the target graph is ob-servable such as public web documents; however, link information in graphs in the real world, such as human relationship or ecoin grapus in the real world, such as human relationship or eco-nomic activities, is rarely open to public. If link analysis can be performed using graphs with private links in a privacy preserving open to send only in compacted with arrivate links of a privacy preserving the send of the send performed using graphs with private titles in a privacy preserving way, it enables us to rank entities connected with private ties, such way, it enables us to rain entities connected with private ites, such as people, organizations, or business transactions. In this paper, as people, organizations, or business transactions. In this paper, we present a secure link analysis for graphs with private links by we present a secure tink analysis for grapus with private tinks by means of cryptographic protocols. Our solutions are designed as a secure tink analysis analysis made. means or cryptograpme protocots. Our solutions are designed as privacy preserving expansions of well-known link analysis methods. ods, PageRank and HITS. The outcomes of our protocols are comous, ragereant and r1115. The outcomes of our protocols are com-pletely equivalent to those of PageRank and HITS. Furthermore, possery equivasem to mose or rageroans and 11113. Furnismore, our protocols theoretically guarantee that the private link information mosessed by each node is not revealed to other nodes. tion possessed by each node is not revealed to other nodes.

Categories and Subject Descriptors H.4 [Information Systems Applications]: Miscellaneous

General Terms

Keywords

link analysis, privacy, ranking, HITS, PageRank

1. INTRODUCTION

Link-based analysis has been developed in the form of algo-Link-based analysis has been developed in the form of algo-rithms that discover useful information from the link structure of minus that discover useria information from the flux structure of the basis basis according to the state of t munually anking entities. In particular, 11/15 [7] and Fagerians.

[9] have been successfully used for the ranking of hyperlinked web by nave oven successionly used for the ranking of hyperinard web documents. These link analysis methods were originally designed for the analysis of make documents. Invasions there can be exactly documents. These link analysis methods were originally designed for the analysis of web documents; however, these can be readily applied to mutually linking entities, such as referenced academic appared to indicately interactions, social as in papers, protein-protein interactions, and so on.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are bear this notice and the full citation on the first page. To copy otherwise, to not made or distributed for profit or commercial accumage and mat copies bear this notice and the full clintion on the first page. To copy otherwise, to receive to not on server or to redistribute to lists, receives nrior specific bear this notice and the ruil cuntion on the first page. To copy otherwise, to requires and/or a fee. permission anaior a fee.

SIGIR '09, July 19-23, 2009, Boston, Massachusetts, USA.

Copyright 2009 ACM 978-1-60558-483-6/09/07 ...55.00.

Shigenobu Kobayashi Tokyo Institute of Technology 4259 Nagatsuta-cho Yokohama, Japan kobayasi@dis.titech.ac.jp

In general, link analysis methods take the entire link structure In general, link analysis methods take the entire link structure as its input. Indeed, for the computation of Google's PageRank as its input. Indeed, for the computation of Google's Pagerana, the linking structures of web documents are collected by crawling the linking structures of web documents are conected by crawing agents which actually wander around public web documents. The same holds for citation graphs of academic papers or interaction

and a second papers of interaction se same nosas for cusmon grapus of academic papers or interaction graphs of protein networks. As shown, existing link analysis meth-Empire of protein networks. As shown, existing tink analysis metroods have inherently been designed based on the fact that the entire ous nave university oven designed based on the fact that the entire link structure of the target graph is observable; however, link informaxion in the real world, such as human relationships or economic activities, is rarely open to public.

In this paper, we present link analysis solutions for graphs of actions of actions and actions are actions as a second action and actions are actions as a second action actions and actions are actions as a second action actions and actions are actions as a second action actions are actions as a second action action actions and actions are actions as a second action action action actions are actions as a second action actions action in this paper, we present this analysis solutions for graphs of privately connected entities. Let there be a directed weighted graph privately connected entities. Let there be a directed weighted graph G = (V, E, W) where V is a set of vertices, E is a set of edges, and W is a second this nature was according that G = (V, E, W) where V is a set of vertices, E is a set of edges, and W is a weight matrix. Throughout this paper, we assume that and W is a weight marrix. Introduction this paper, we assume that the set of vertices corresponds to a collection of distributed nodes where the communicational volume of made mode is real-married. Educathe set of vertices corresponds to a collection of distributed nodes where the computational power of each node is polynomial. Edges where the computational power of each node is polynomias, sugar correspond to links between nodes; weights of edges correspond to weights of these links. Let there be a link of node i pointing to node weigns of these mass, Let mere be a mix of more i pointing to more

j. In our setting, we assume that link eij and weight of the link with

and more deciral to be known by modes when these mode i and mode. J. In our sessing, we assume that anx ej and weight of the anx wij are not desired to be known by nodes other than node i and node i Eusthamour was during our link analysis evaluations beauted on the are not desired to be known by nodes other than node i and node j. Furthermore, we design our link analysis solutions based on the three privacy models of graphs described as below:

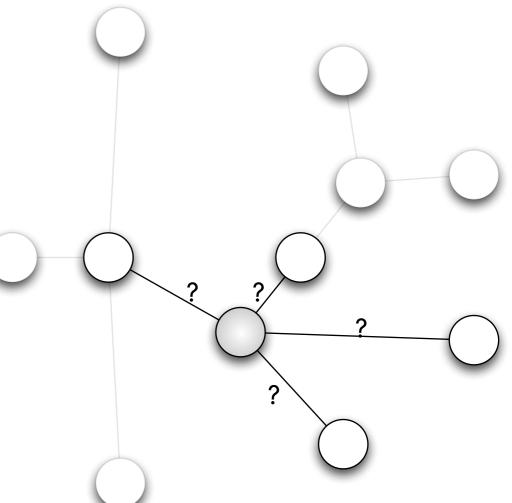
Weight-aware model. If both the head node i and the tail node Weight-aware more, it out the near none rains the san more

j know the existence of the link and the weight value, this is des-J know the existence of the tink and the weight value, this is uestigated as weight-aware link-aware model (or weight-aware model ignated as weight-aware time-aware model (or weight-aware model for short). For example, consider commercial relationships among transactions transactions made transactions with enterprises. Each enterprise may conduct business transactions with emerprises, each emerprise may commer oursiness consocious with the other enterprises. Let the *i*th enterprise purchase some products constitute the management of the bounded by the bounded the other enterprises. Let the rin enterprise purchase some products from the jth enterprise. This transaction corresponds to link e_{ij} and from the Jin emergerse. This transaction corresponds to ank e/i and the transaction value corresponds to weight w/j. In this case, both the objective one success of the objective order. the transaction value corresponds to weight w_{ij} . In this case, both the ith and jth enterprise are aware of the existence of this link and

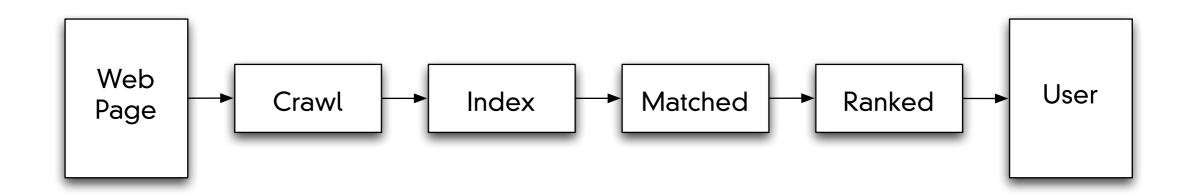
the tri and fin emerprise are aware of the existence of this time and know the weight value, but enterprises other than i and j do not be existenced of this transaction and the transaction value. know the weight value, our emerprises other than 1 and 1 oo n know the existence of this transaction and the transaction value. Link-aware model. If the head node i and the tail node j know Link-aware modes, it the near node I and the tan node I know the existence of the link, but the weight value is only known by the the extractive or tree time, one me weight verse is only known by the head node i, this is designated as line aware weight-unaware model for charge. For assumation remains and time of head node i, this is designated as time-aware weight-anaware model (or link-aware model for short). For example, consider call logs of the consider call logs of the consider i. This call (or tink-aware model for short). For example, consider call logs of cell-phones. Let caller i make a phone call to receiver j. This call corresponds to link eq and the probability that i makes a phone call to this count book on the sociality of a this count book on the sociality of a star country. corresponds to the weight w_{ij} of e_{ij} . In this case, both caller is and receiver j are aware of the existence of the link, but the call probability will are known only by caller i.

Link-unaware model. If only the head node i knows the exis-Link-unaware model. It only the nead node I knows the exis-tence of the link and the weight value, but the tail node J knows nothing, this is designated as link-unaware weight-unaware model nothing, this is designated as time-unaware weight-unaware model (or link-unaware model for short). For example, consider a peer (or max-massace mones are season, ever example, commences of choice a limited number of other mondates. Each in evanuation science unoug memoris or person can choose a limited number of other membranes.

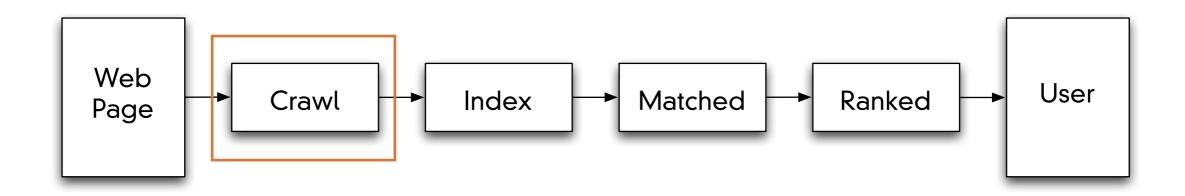
- Crawl Selection Policy
 - A good page, not crawled, is not returned for a query
 - Deciding what to crawl can make a difference with finite resources



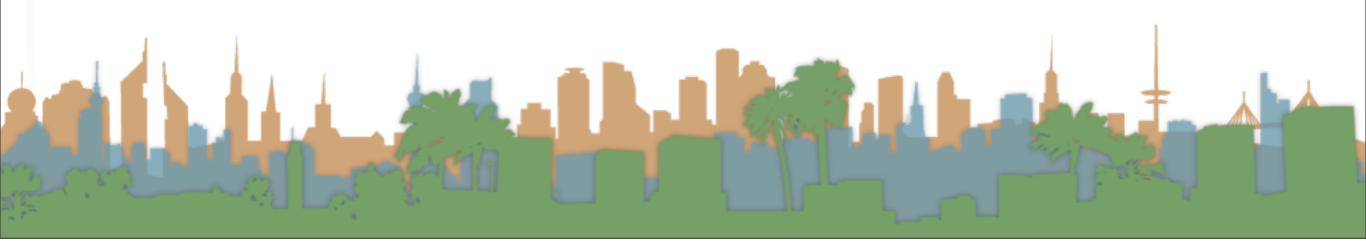


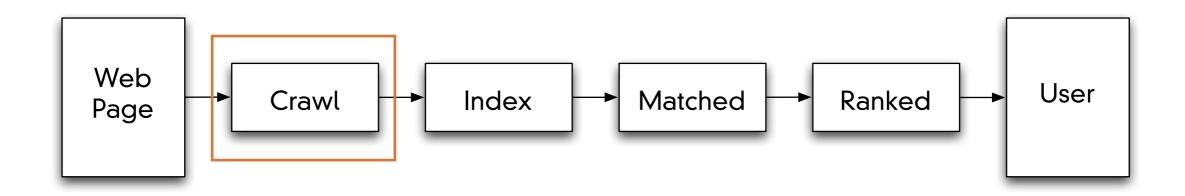


- For a web page to get to a user in a search result many things must go well
- A failure results in worse quality and unhappy users
- Instead of holding the corpus constant and testing retrieval algorithms, this paper varies the corpus through crawl strategies



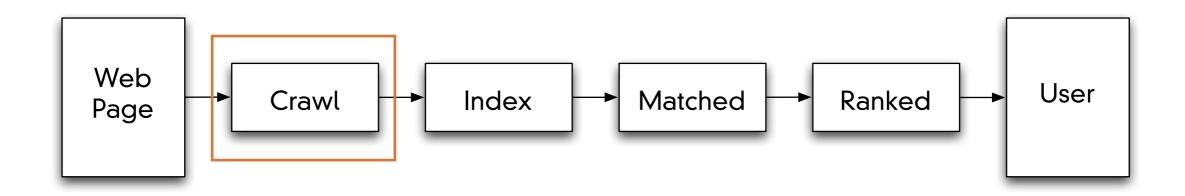
- Successful crawling depends on
 - size of crawl
 - crawl selection policy





- Improving crawling depends on
 - Increasing the size of the crawl, or
 - using a better crawl selection policy





- Improving crawling depends on
 - Increasing the size of the crawl, or
 - using a better crawl selection policy



- This paper
 - Provides a metric to compare crawls
 - evaluates 4 different crawl policies
 - one-time
 - incrementally



- A naive approach to crawling would be to start from a seed set every time
- A smarter approach is to crawl based on quality metrics derived from the last crawl



- 4 crawl selection policies
 - breadth first search (baseline)
 - indegree
 - trans-domain indegree
 - highest PageRank
- Not done:
 - focussing crawl based on queries with lame results



- Options for evaluating a crawl
 - Efficiency
 - politeness
 - queueing
 - data structures
 - resource allocation
 - "Goodness" of collected corpus



- Goodness can be measure by "RankMass"
 - Like precision over PageRank
 - What % of top N pages did you crawl?
 - Requires a big crawl for ground truth
 - Requires crawler
- Alternative is Normalized Discounted Cumulative
 Gain (NDCG)



- metric: maxNDCG
 - eliminates variation based on retrieval algorithm
 - several thousand test queries
 - 10,570 sampled queries (Microsoft)
 - maxNDCG = "how many of the web pages returned by 3 large web search engines are in your crawl set?" weighted by human judged relevance



- metric: NDCG@K
 - G(j) is user rating of document at position j

$$\sum_{j=1}^{K} \left(\frac{G(j)}{\log(1+j)} \right)$$



- metric: click utility
 - sum of clicks on pages in crawled corpus
 - requires search engine support (Microsoft)
 - not weighted by rank in query results
 - (multiple queries)
 - like RankMass for clicks

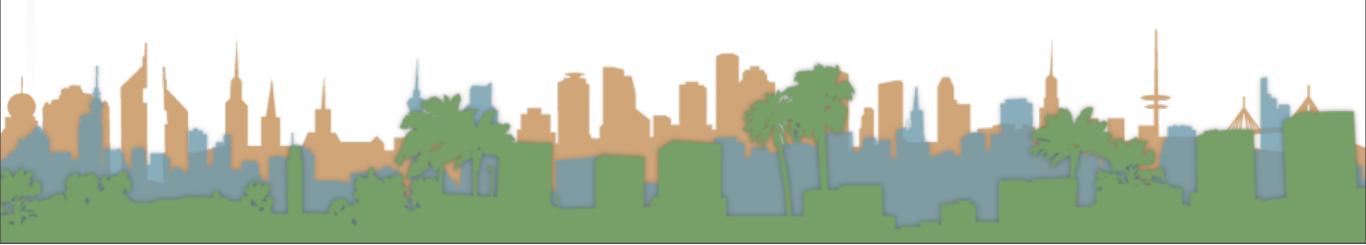


- incremental metric: Jaccard coefficent
 - High Jaccard coefficient means not many pages changed between crawls

$$\frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$



- incremental metric: churn
 - How many pages came and went during the iterative crawls
 - Users care because it makes the search engine look stable



- Systems issues
 - Hard to compare crawls when the Internet is dynamic
 - Real-time events are irreproducible
 - network transience
 - For this experiment
 - Sandbox crawl (basically a big cache)
 - First time a URL is requested it is cached.
 - Time is not being evaluated
 - When required equal scores are randomly selected
 - for a ranked list (like trans-domain indegree)

- Experiments
 - Baseline
 - breadth-first crawl starting at open-directory
 - 930,320,010 filtered pages recovered

Arts	<u>Business</u>	<u>Computers</u>
Movies, Television, Music	Jobs, Real Estate, Investing	Internet, Software, Hardware

Games Health Home

Video Games, RPGs, Gambling... Fitness, Medicine, Alternative... Family, Consumers, Cooking...

Kids and Teens News Recreation

Arts, School Time, Teen Life... Media, Newspapers, Weather... Travel, Food, Outdoors, Humor...

Reference Regional Science

Maps, Education, Libraries... US, Canada, UK, Europe... Biology, Psychology, Physics...

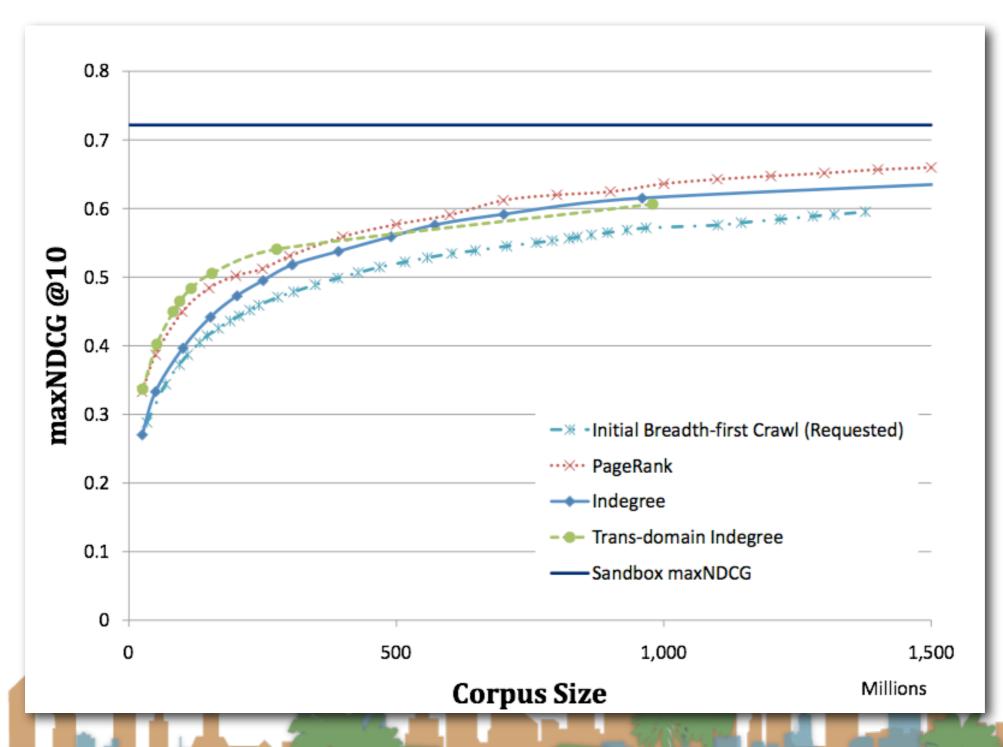
Shopping Society Sports

Clothing, Food, Gifts... People, Religion, Issues... Baseball, Soccer, Basketball...

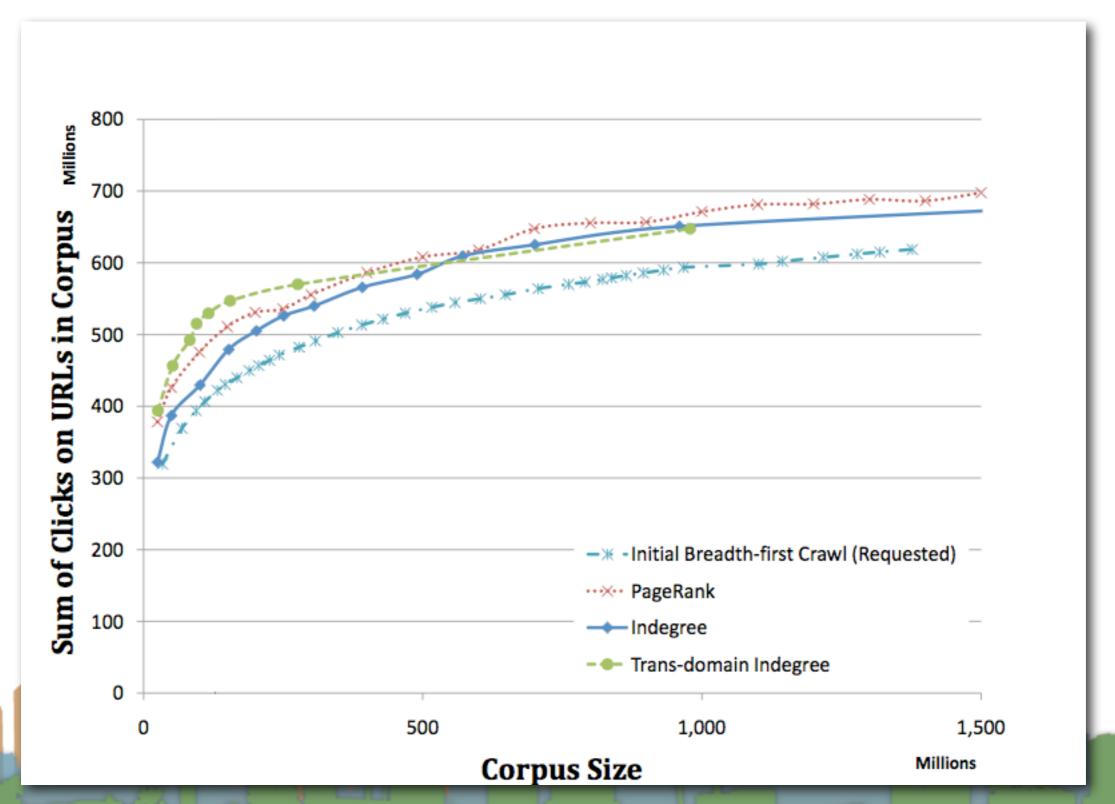
World

Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

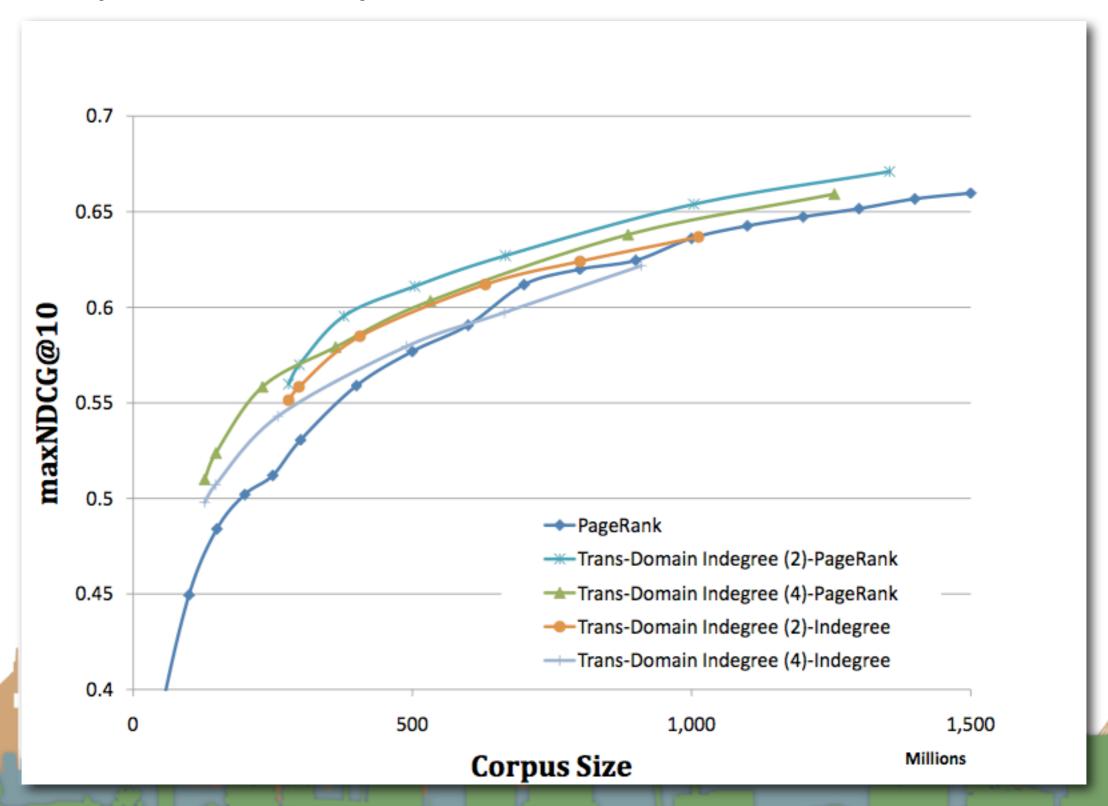
Experiments - Single Crawl



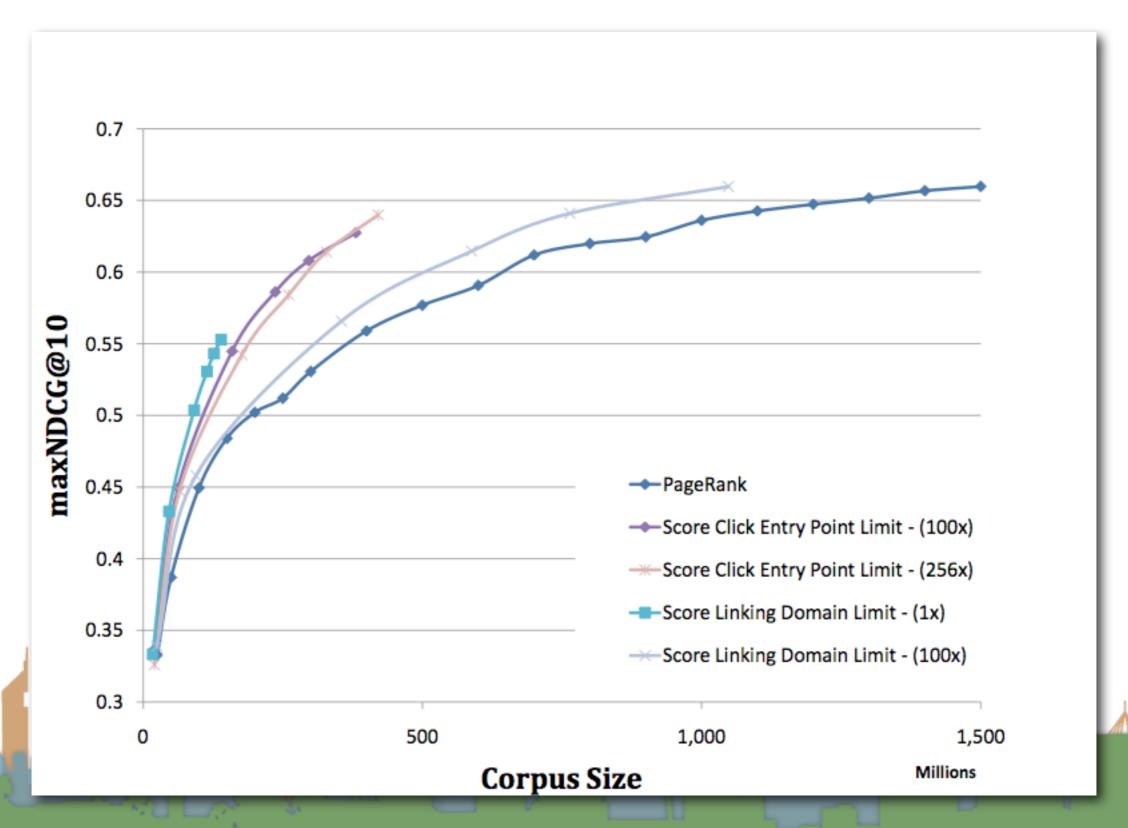
Experiments - Single Crawl



Experiments - Hybrid Crawl



Experiments - Domain Limits on PageRank Crawl



Experiments - Iterative Behavior

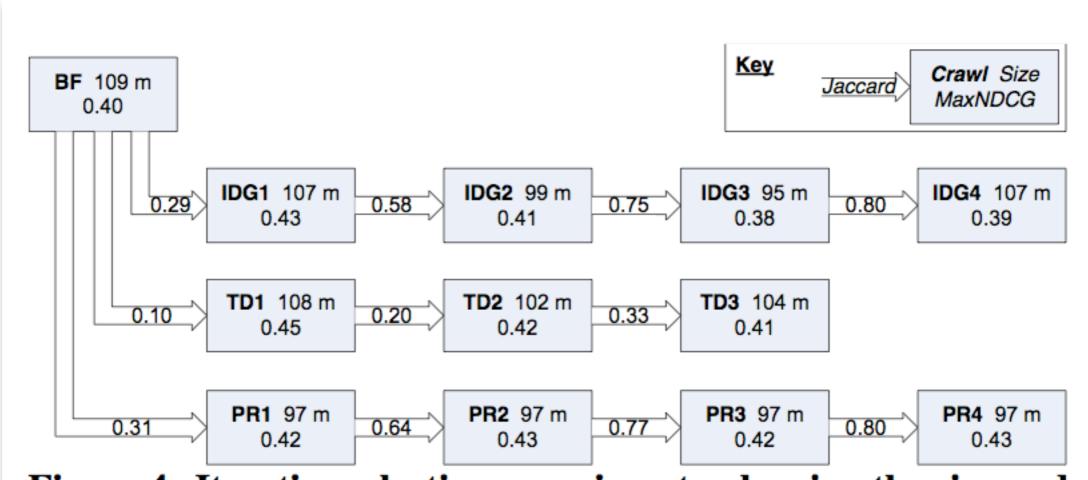


Figure 4: Iterative selection experiments, showing the size and maxNDCG of requested selections, and the Jaccard similarity between iterations.

Experiments - Iterative Set Similarity

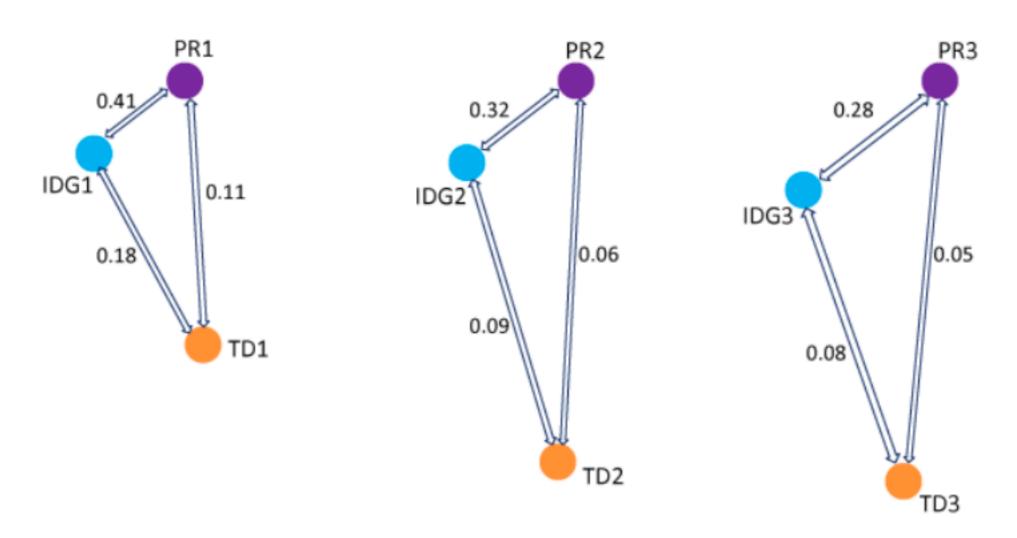
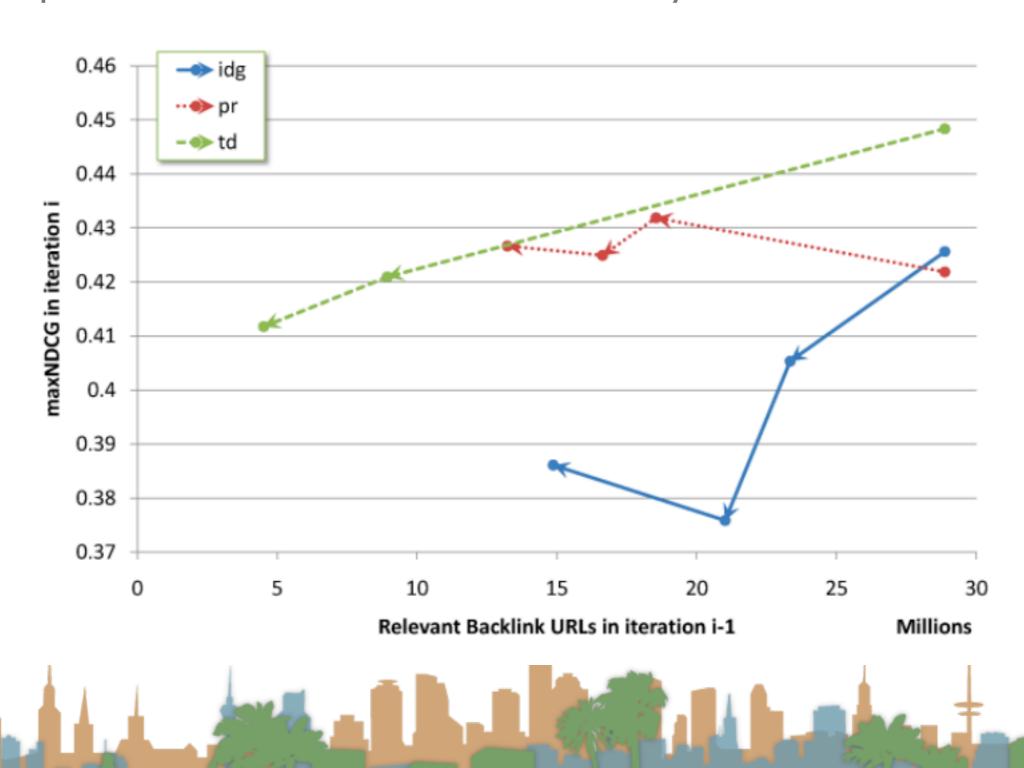


Figure 5: Jaccard similarity between corpora selected by different policies at different iterations.

Experiments - Iterative Self-Similarity



Experiments - Iterative Stability

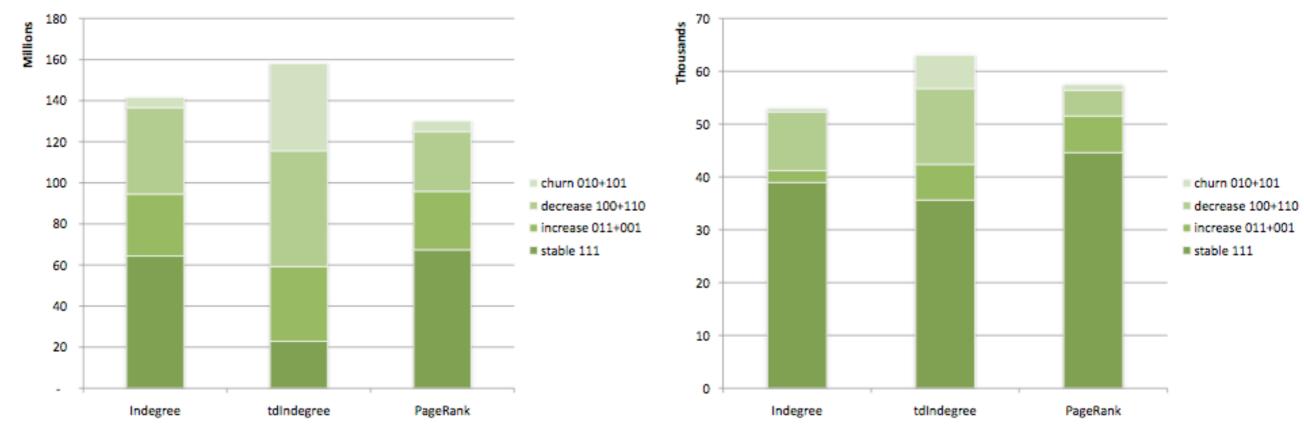


Figure 6: Stability of crawl policies, over three iterations. Left: All URLs. Right: Relevant URLs



- Take Away
 - Use a combination of TD and PageRank to guide crawl
 - Don't overcrawl one domain
 - maxNDCG is a measure comparable to user click utility



Advances in Link Analysis

"The Impact of Crawl Policy on Web Search Effectiveness" by Fetterly, Craswell, Vinay SIGIR2009

"Link Analysis for Private Weighted Graphs" by Sakuma, Kobayashi SIGIR2009

The Impact of Crawl Policy on Web Search Effectiveness

Dennis Fetterly Microsoft Research Mountain View, CA USA fetterly@microsoft.com

Nick Craswell Microsoft Research Cambridge, UK nickcr@microsoft.com

Cambridge, UK vvinay@microsoft.com

Crawl selection policy has a direct influence on Web search effectiveness, because a useful page that is not selected for crawing will also be absent from course results. Yet shows here there are the course to be absent from course process. also be absent from search results. Yet there has been little or no work on measuring this effect. We introduce an evaluation framework, based on relevance judgments pooled from multiple search engines, measuring the maximum potential NDCG that is achievable engines, measuring the maximum potential NDCG that is achievable using a particular crawl. This allows us to evaluate different crawl policies and investigate important scenarios like selection stability poncies and investigate important scenarios like selection stability over multiple iterations. We conduct two sets of crawling experiover multiple iterations. We conduct two sets of crawing experi-ments at the scale of 1 billion and 100 million pages respectively. ments at the scare of 1 minor and 100 minor pages respectively.

These show that crawl selection based on PageRank, indegree and trans-domain indegree all allow better retrieval effectiveness than a simple breadth-first crawl of the same size. PageRank is the most reliable and effective method. Trans-domain indegree can outperform PageRank, but over multiple crawl iterations it is less effective and regenerity, our ever manager craws measures it is less effective and more unstable. Finally we experiment with combinations of crawl and the combination of crawless and the combination of crawless and the combination of crawless and combination of crawle more unstable. Finally we experiment with combinations of crawl selection methods and per-domain page limits, which yield crawls with greater potential NDCG than PageRank.

Categories and Subject Descriptors H.3.3 [Information Storage and Retrieval]: Information Search

and Retrieval

General Terms

Measurement, Experimentation

A useful Web search result will only be seen by users if it is 1. INTRODUCTION erawled by the search engine, indexed correctly, found in the index crawieu by the search engine, indexed correctly, iound in the index when matched with a query and ranked highly in the search result. listing. It only takes one failure in this chain of events for the useful issing. It only taxes one tanire in his crain of events for the userial (relevant) result to be lost. If such failures happen often, users

will perceive a drop in the quality of search results. Therefore, to will perceive a crop in the quarry of search results. Therefore, to optimize user satisfaction, it is important to avoid failure at every stage.

Success at the crawling stage depends on the size of the crawl and the crawl selection policy. For example, the policy of preferring

Permission to make digital or hard copies of all or part of this work for personnal or electronic tree is presented without fee perceided that coming are Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to bear this notice and the full citation or to redistribute to lists, requires prior specific

pages with highest PageRank [7] and a size limit of N leads to pages with nighest PageRank 1/1 and a size firmt of N leads to selecting a set of N high-PageRank pages. When searches are carried out, the quality of search results will sometimes be reduced carried out, the quality of search results will sometimes be reduced because pages that would have been relevant and retrievable were not selected for crawling. One way to reduce such failures is to not selected for crawling. One way to reduce such failures is to increase the size N of the crawl. Another approach is to improve

Although well-known methods exist for evaluating search rel-Annough well-known memous exist for evaruating search reference, such as NDCG [13], we are not aware of any published evance, such as NLCAU [13], we are not aware or any punnsned experiments that compare the relevance achievable by different experiments that compare the relevance acmevante by different crawl policies. Acting as a barrier to experimentation are the large communication and computational costs of conducting multiple communication and computational costs of considering natural crawls, creating multiple indices and processing queries. Our framework appealment of the plan appeal appealment and an appealment of the plan appealment and appealment appealment appealment and appealment appealme work ameliorates this via a crawl sandbox and an evaluation metric work amenorates this via a crawi santonox and an evaluation metre that only requires the set of selected URLs. The sandbox is simply apply to the control of a cache, to avoid crawling URLs more than once if selected by multiple policies or iterations. The metric, maxNDCG, is the best numple policies or nerations. The meme, many code, is the seed potential NDCG that could be achieved based on the presence or absence of relevant pages in a crawl, maxNDCG is proportional to NDCG but may be calculated without indexing and retrieval. It may NIA-U out may or carcurated without indexing and redieval. It may even be calculated for a selected set of pages without attempting to even or carcurated for a selected set of pages without attempting a crawl them, estimating the NDCG that would be achievable by perfect ranker if all selected pages were successfully crawled. These efficiency techniques allow us to run a large numb

experiments comparing crawl policies. We focus on policy experiments comparing craws principes. We tocal on principles selecting a new crawl based on the link graph of a previous cr Seaccung a new craws observed in one and grape of a previous ca This is a common scenario, allowing an engine to shift its s towards pages that are preferred according to some link-bay (such as PageRank) but not yet included in the crawl. In Section 2, we discuss different aspects involved in

in occuon 2, we ensure uniferent aspects involved in tion of crawling methods. We provide motivation for or tal setup, and where appropriate, we provide referen work. We then present experiments in Section 3 and

2. CRAWLING AND EVALUAT Search engines are the primary discovery to on the Web, and the Web has an effectively on use wee, and the wee has an effectively that might be indexed. A search engine mus pages it indexes, to make the best use of its Search engines use crawlers to download which indices are built. Starting from a download pages and extract links. It rec seen and adds any unseen URLs to a s After an initial crawl, there is the

corpus, since pages are continually deleted [4]. One option would be to although completely throwing ay

Link Analysis for Private Weighted Graphs

University of Tsukuba 1-1-1 Tennodai, jun@cs.tsukuba.ac.jp

ABSTRACT

Link analysis methods have been used successfully for knowledge Little analysis memous nave been used successionly for knowledge discovery from the link structure of mutually linking entities. Exdiscovery from the link structure of mutually linking entities. Ex-isting link analysis methods have been inherently designed based on the fact that the entire link structure of the target graph is obon the fact that the entire and structure of the target graph is ob-servable such as public web documents; however, link information in graphs in the real world, such as human relationship or ecoin grapus in the real world, such as human relationship or eco-nomic activities, is rarely open to public. If link analysis can be performed using graphs with private links in a privacy preserving open to send only in compacted with arrivate links of a privacy preserving the send of the send performed using graphs with private titles in a privacy preserving way, it enables us to rank entities connected with private ties, such way, it enables us to rain entities connected with private ites, such as people, organizations, or business transactions. In this paper, as people, organizations, or business transactions. In this paper, we present a secure link analysis for graphs with private links by we present a secure tink analysis for grapus with private tinks by means of cryptographic protocols. Our solutions are designed as a secure tink analysis analysis made. means or cryptograpme protocots. Our solutions are designed as privacy preserving expansions of well-known link analysis methods. ods, PageRank and HITS. The outcomes of our protocols are comous, ragereant and r1115. The outcomes of our protocols are com-pletely equivalent to those of PageRank and HITS. Furthermore, possery equivasem to mose or rageroans and 11113. Furnismore, our protocols theoretically guarantee that the private link information mosessed by each node is not revealed to other nodes. tion possessed by each node is not revealed to other nodes.

Categories and Subject Descriptors H.4 [Information Systems Applications]: Miscellaneous

General Terms

Keywords

link analysis, privacy, ranking, HITS, PageRank

1. INTRODUCTION

Link-based analysis has been developed in the form of algo-Link-based analysis has been developed in the form of algo-rithms that discover useful information from the link structure of minus that discover useria information from the flux structure of the basis basis according to the state of t munually anking entities. In particular, 11/15 [7] and Pagerlank
[9] have been successfully used for the ranking of hyperlinked web by nave oven successionly used for the ranking of hyperinard web documents. These link analysis methods were originally designed for the analysis of make documents. Invasions there can be exactly documents. These link analysis methods were originally designed for the analysis of web documents; however, these can be readily applied to mutually linking entities, such as referenced academic appared to indicately interactions, social as in papers, protein-protein interactions, and so on.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are bear this notice and the full citation on the first page. To copy otherwise, to not made or distributed for profit or commercial accumage and mat copies bear this notice and the full clintion on the first page. To copy otherwise, to receive to not on server or to redistribute to lists, receives nrior specific bear this notice and the ruil cuntion on the first page. To copy otherwise, to requires and/or a fee. permission anaior a fee.

SIGIR '09, July 19-23, 2009, Boston, Massachusetts, USA.

Copyright 2009 ACM 978-1-60558-483-6/09/07 ...55.00.

Shigenobu Kobayashi Tokyo Institute of Technology 4259 Nagatsuta-cho Yokohama, Japan kobayasi@dis.titech.ac.jp

In general, link analysis methods take the entire link structure In general, link analysis methods take the entire link structure as its input. Indeed, for the computation of Google's PageRank as its input. Indeed, for the computation of Google's Pagerana, the linking structures of web documents are collected by crawling the linking structures of web documents are conected by crawing agents which actually wander around public web documents. The same holds for citation graphs of academic papers or interaction

and a second papers of interaction se same nosas for cusmon grapus of academic papers or interaction graphs of protein networks. As shown, existing link analysis meth-Empire of protein networks. As shown, existing tink analysis metroods have inherently been designed based on the fact that the entire ous nave university oven designed based on the fact that the entire link structure of the target graph is observable; however, link informaxion in the real world, such as human relationships or economic activities, is rarely open to public.

In this paper, we present link analysis solutions for graphs of actions of actions and actions are actions as a second action and actions are actions as a second action actions and actions are actions as a second action actions and actions are actions as a second action actions are actions as a second action action actions and actions are actions as a second action action action actions are actions as a second action actions action in this paper, we present this analysis solutions for graphs of privately connected entities. Let there be a directed weighted graph privately connected entities. Let there be a directed weighted graph G = (V, E, W) where V is a set of vertices, E is a set of edges, and W is a second this nature was according that G = (V, E, W) where V is a set of vertices, E is a set of edges, and W is a weight matrix. Throughout this paper, we assume that and W is a weight marrix. Introduction this paper, we assume that the set of vertices corresponds to a collection of distributed nodes where the communicational volume of made mode is real-married. Educathe set of vertices corresponds to a collection of distributed nodes where the computational power of each node is polynomial. Edges where the computational power of each node is polynomias, sugar correspond to links between nodes; weights of edges correspond to weights of these links. Let there be a link of node i pointing to node weigns of these mass, Let mere be a mix of more i pointing to more

j. In our setting, we assume that link eij and weight of the link with

and more deciral to be known by modes when these mode i and mode. J. In our sessing, we assume that anx ej and weight of the anx wij are not desired to be known by nodes other than node i and node i Eusthoanness was during our link analysis evaluations beautiful. are not desired to be known by nodes other than node i and node j. Furthermore, we design our link analysis solutions based on the three privacy models of graphs described as below:

Weight-aware model. If both the head node i and the tail node Weight-aware more, it out the near none rains the san more

j know the existence of the link and the weight value, this is des-J know the existence of the tink and the weight value, this is uestigated as weight-aware link-aware model (or weight-aware model ignated as weight-aware time-aware model (or weight-aware model for short). For example, consider commercial relationships among transactions transactions made transactions with enterprises. Each enterprise may conduct business transactions with emerprises, each emerprise may commer oursiness consocious with the other enterprises. Let the *i*th enterprise purchase some products constitute the management of the bounded by the bounded the other enterprises. Let the rin enterprise purchase some products from the jth enterprise. This transaction corresponds to link e_{ij} and from the Jin emergerse. This transaction corresponds to ank e/i and the transaction value corresponds to weight w/j. In this case, both the objective one success of the objective order. the transaction value corresponds to weight w_{ij} . In this case, both the ith and jth enterprise are aware of the existence of this link and

the tri and fin emerprise are aware of the existence of this time and know the weight value, but enterprises other than i and j do not the existences of this transaction and the transaction value know the weight value, our emerprises other than 1 and 1 oo n know the existence of this transaction and the transaction value. Link-aware model. If the head node i and the tail node j know Link-aware modes, it the near node I and the tan node I know the existence of the link, but the weight value is only known by the the extractive or tree time, one me weight varies is only known by the head node i, this is designated as line aware weight-unaware model for charge. For assumation remaining out time of head node i, this is designated as time-aware weight-anaware model (or link-aware model for short). For example, consider call logs of the consider call logs of the consider i. This call (or tink-aware model for short). For example, consider call logs of cell-phones. Let caller i make a phone call to receiver j. This call corresponds to link eq and the probability that i makes a phone call to this count book on the sociality of a this count book on the sociality of a star country. corresponds to the weight w_{ij} of e_{ij} . In this case, both caller is and receiver j are aware of the existence of the link, but the call probability will are known only by caller i.

Link-unaware model. If only the head node i knows the exis-Link-unaware model. It only the nead node I knows the exis-tence of the link and the weight value, but the tail node J knows nothing, this is designated as link-unaware weight-unaware model nothing, this is designated as time-unaware weight-unaware model (or link-unaware model for short). For example, consider a peer (or max-massace mones are season, ever example, commences of choice a limited number of other mondates. Each in evanuation science unoug memoris or person can choose a limited number of other membranes.