Is Wikipedia Link Structure Different?

Jaap Kamps, Marijn Koolen
University of Amsterdam

WSDM'09
Barcelona, 12 February, 2009

Motivation

- The links in Wikipedia are a special case of the general hyperlinks that connect the Web:
  - ⋆ Links signal semantic relation, not only serve navigational purposes.

Motivation

- The links in Wikipedia are a special case of the general hyperlinks that connect the Web:
  - ⋆ Links signal semantic relation, not only serve navigational purposes.
- http://en.wikipedia.org/wiki/Wikipedia:Only_make_links_that_are_relevant_to_the_context

Motivation

- The links in Wikipedia are a special case of the general hyperlinks that connect the Web:
  - ⋆ Links signal semantic relation, not only serve navigational purposes.
- http://en.wikipedia.org/wiki/Wikipedia:Only_make_links_that_are_relevant_to_the_context
- Is Wikipedia link structure different?

Outline

- Introduction
- Comparative Analysis of Link Structure
  - ⋆ Are there differences between degree distributions of incoming and outgoing links?
  - ⋆ How does the link topology relate to relevance of retrieval results?
- Link Evidence in Retrieval
  - ⋆ What is the impact of link evidence on effectiveness of Wikipedia and Web retrieval?
- Conclusions

Introduction

- Link evidence has been exploited to improve information retrieval on the web
  - ⋆ PageRank [Page et al., 1998] exploits global web structure
  - ⋆ and HITS [Kleinberg, 1999] exploits local web structure.

Introduction

- Link evidence has been exploited to improve information retrieval on the web
  - ⋆ PageRank [Page et al., 1998] exploits global web structure
  - ⋆ and HITS [Kleinberg, 1999] exploits local web structure.
- Commercial search engine companies have heralded the use of link structure as one of their key technologies.

Introduction

- Link evidence has been exploited to improve information retrieval on the web
  - ⋆ PageRank [Page et al., 1998] exploits global web structure
  - ⋆ and HITS [Kleinberg, 1999] exploits local web structure.
- Commercial search engine companies have heralded the use of link structure as one of their key technologies.
- TREC experiments failed to establish the effectiveness of link evidence for general ad hoc retrieval on Web collections
  - ⋆ this lead to the introduction of Web-centric tasks like homepage, topic distillation, etc.

Wikipedia Specific Aspects

- Some aspects specific to Wikipedia might affect the nature of links

Wikipedia Specific Aspects

- Some aspects specific to Wikipedia might affect the nature of links
- Due to encyclopedic organisation of Wikipedia:
  - ⋆ it is clear what information is there
  - ⋆ there is low redundancy of information
  - ⋆ therefore, it is clear where to link to

Wikipedia Specific Aspects

- Some aspects specific to Wikipedia might affect the nature of links
- Due to encyclopedic organisation of Wikipedia:
  - ⋆ it is clear what information is there
  - ⋆ there is low redundancy of information
  - ⋆ therefore, it is clear where to link to
- Due to shared authorship:
  - ⋆ "missing" links are added by others or link bots
  - ⋆ they can also remove uninformative links

Wikipedia Specific Aspects

- Some aspects specific to Wikipedia might affect the nature of links
- Due to encyclopedic organisation of Wikipedia:
  - ⋆ it is clear what information is there
  - ⋆ there is low redundancy of information
  - ⋆ therefore, it is clear where to link to
- Due to shared authorship:
  - ⋆ "missing" links are added by others or link bots
  - ⋆ they can also remove uninformative links
- Is Wikipedia link structure different?

Outline

- Introduction
- **Comparative Analysis of Link Structure**
  - ⋆ **Are there differences between degree distributions of incoming and outgoing links?**
  - ⋆ **How does the link topology relate to relevance of retrieval results?**
- Link Evidence in Retrieval
  - ⋆ What is the impact of link evidence on effectiveness of Wikipedia and Web retrieval?
- Conclusions

Web and Wikipedia Collections

- Analysis based on IR test-collections:
  - ⋆ TREC Webtrack collection (.GOV, 2002) with 1.2 million documents, 225 topics (known-item and topic distilation)
  - ⋆ INEX 2006 Wikipedia collection with over 650,000 document, 217 ad hoc topics of Ad hoc tracks of 2006 and 2007,

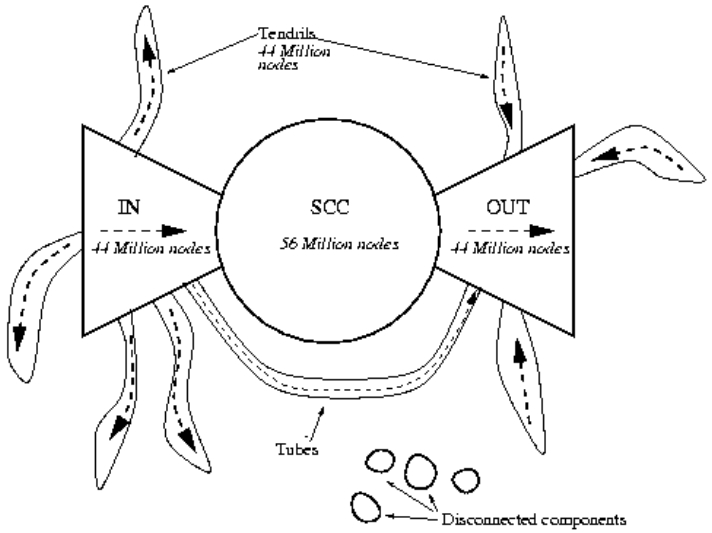
Web and Wikipedia Collections

- Analysis based on IR test-collections:
  - ⋆ TREC Webtrack collection (.GOV, 2002) with 1.2 million documents, 225 topics (known-item and topic distilation)
  - ⋆ INEX 2006 Wikipedia collection with over 650,000 document, 217 ad hoc topics of Ad hoc tracks of 2006 and 2007,
- Is .GOV representative of the Web at large?

Web and Wikipedia Collections

- Analysis based on IR test-collections:
  - ⋆ TREC Webtrack collection (.GOV, 2002) with 1.2 million documents, 225 topics (known-item and topic distilation)
  - ⋆ INEX 2006 Wikipedia collection with over 650,000 document, 217 ad hoc topics of Ad hoc tracks of 2006 and 2007,
- Is .GOV representative of the Web at large?
  - ⋆ Web is infinitely large and heterogeneous
  - ⋆ .GOV collection is a small crawl of a specific domain
  - ⋆ but we expect it to be a good enough approximation for our purposes

Graph Structure of the Web



- Broder et al. [2000] propose a bowtie model of the Web
  - ⋆ Strongly Connected Component (SCC) of 56M pages (28%)
  - ⋆ a set IN containing pages with a path to all SCC
  - ⋆ a set OUT containing pages with a path from all SCC
  - ⋆ Weakly Connected Component (WCC) contains over 90% of the pages

.GOV and Wikipedia Link Topology

- If we look at connectedness:
  - ⋆ .GOV SCC contains 73.16%, WCC contains 96.92%
  - ⋆ Wikipedia SCC contains 91.91%, WCC contains 99.74%

.GOV and Wikipedia Link Topology

- If we look at connectedness:
  - ⋆ .GOV SCC contains 73.16%, WCC contains 96.92%
  - ⋆ Wikipedia SCC contains 91.91%, WCC contains 99.74%
- We look at indegree/outdegree (number of incoming/outgoing links of document):

.GOV and Wikipedia Link Topology

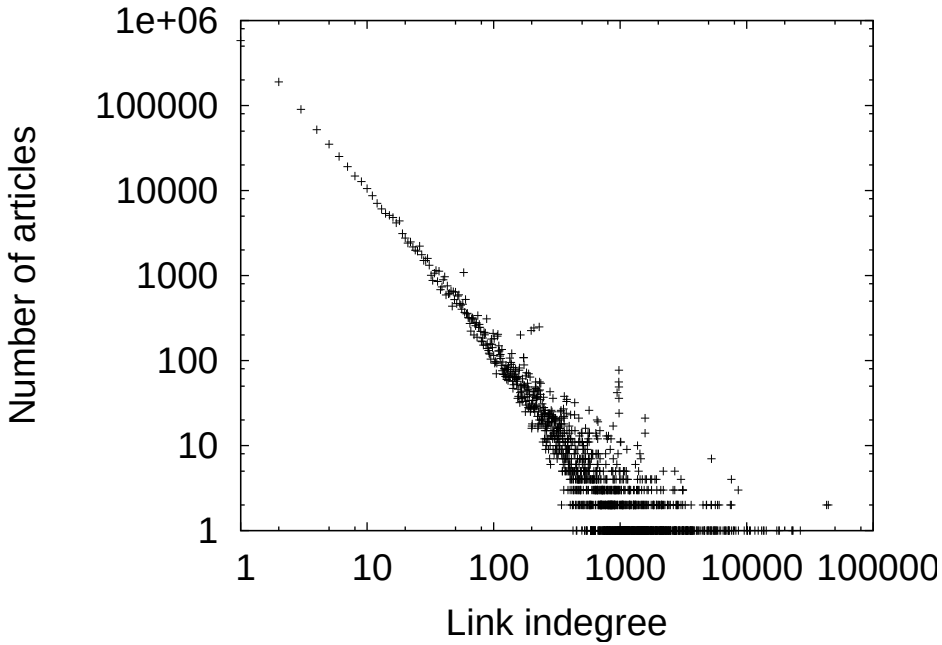
- If we look at connectedness:
  - ⋆ .GOV SCC contains 73.16%, WCC contains 96.92%
  - ⋆ Wikipedia SCC contains 91.91%, WCC contains 99.74%

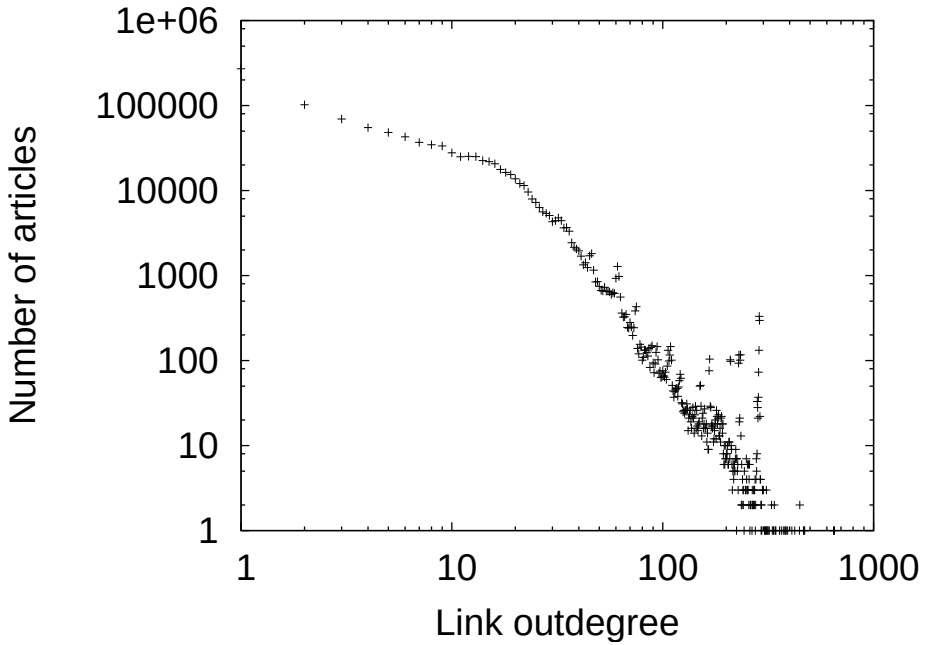
- We look at indegree/outdegree (number of incoming/outgoing links of document):

Collection	degree	min	max	**mean**	median	stdev
.GOV	Indegree	0	44,228	**8.90**	1	126.00
	Outdegree	0	653	**8.90**	4	16.61
Wikipedia	Indegree	0	74,937	**20.63**	4	282.94
	Outdegree	0	5,098	**20.63**	12	36.70

  - ⋆ the average number of (incoming, outgoing) links is 8.90 for .GOV and 20.63 for Wikipedia

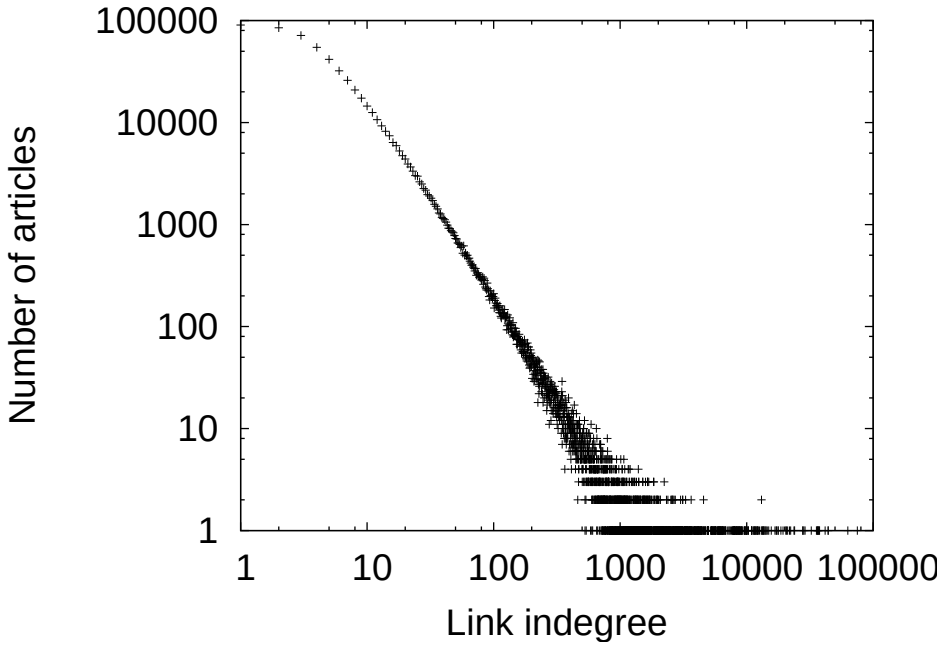
.GOV Degree Distribution

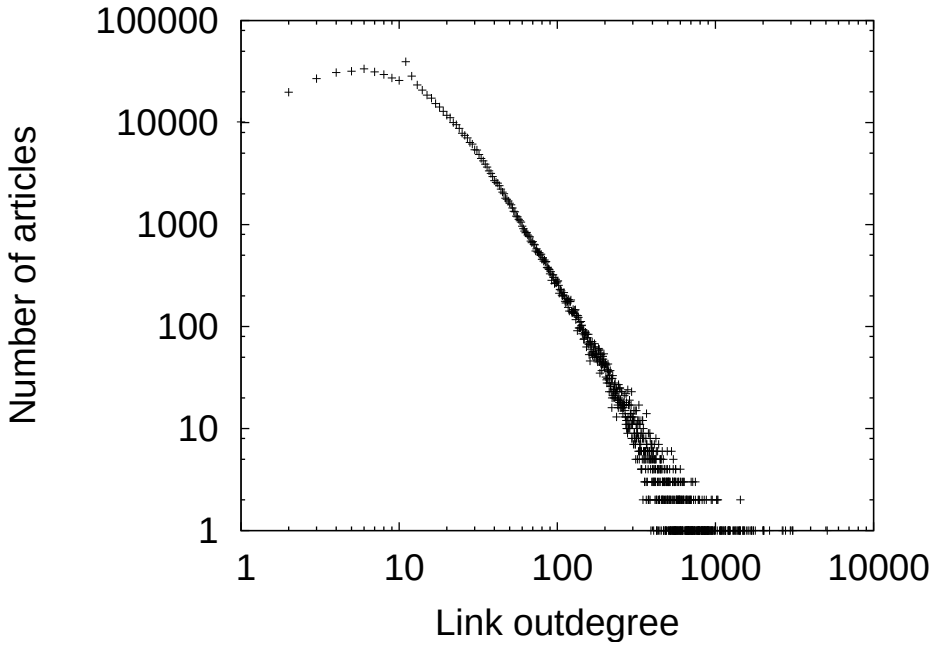




- Outdegree starts less steep:
  - ⋆ Median indegree is 1, median outdegree is 4

Wikipedia Degree Distribution





- Again, outdegree starts less steep:
  - ⋆ Median indegree is 4, median outdegree is 12

Wrap Up Degree Distribution

- Are there differences between degree distributions of incoming and outgoing links?
  - ⋆ both link structures are typical scale-free networks with powerlaw distributions of link degrees, and a single giant connected component.
  - ⋆ Wikipedia is more connected than .GOV
  - ⋆ Wikipedia is more densely linked than .GOV

Wrap Up Degree Distribution

- Are there differences between degree distributions of incoming and outgoing links?
  - ⋆ both link structures are typical scale-free networks with powerlaw distributions of link degrees, and a single giant connected component.
  - ⋆ Wikipedia is more connected than .GOV
  - ⋆ Wikipedia is more densely linked than .GOV
- How does the link topology relate to relevance of retrieval results?
  - ⋆ compute the probability of relevance over degree

Computing Probability of Relevance

- How can we compute the probability of relevance (PoR) over link degrees?

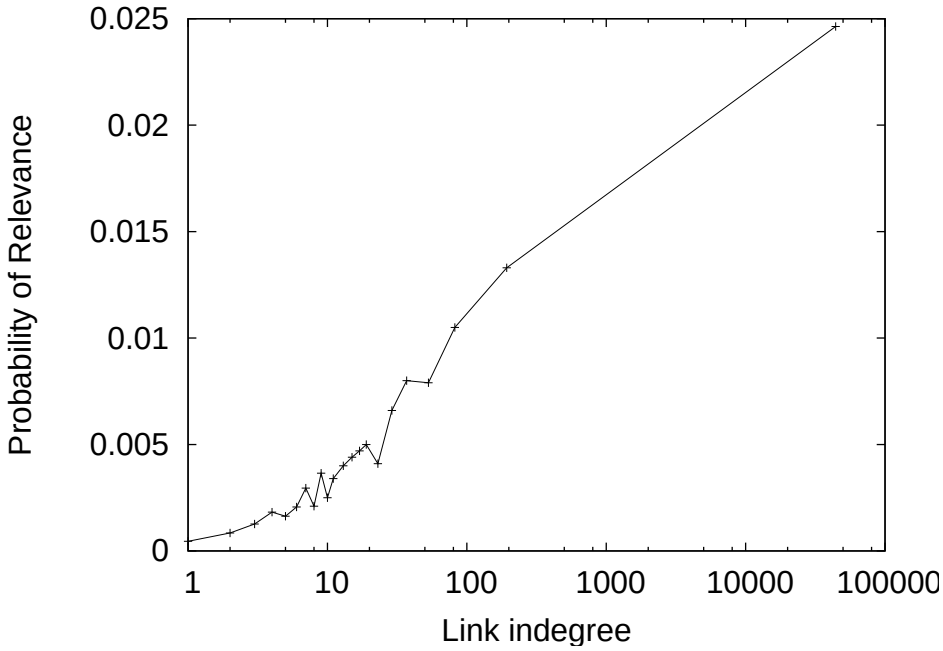
Computing Probability of Relevance

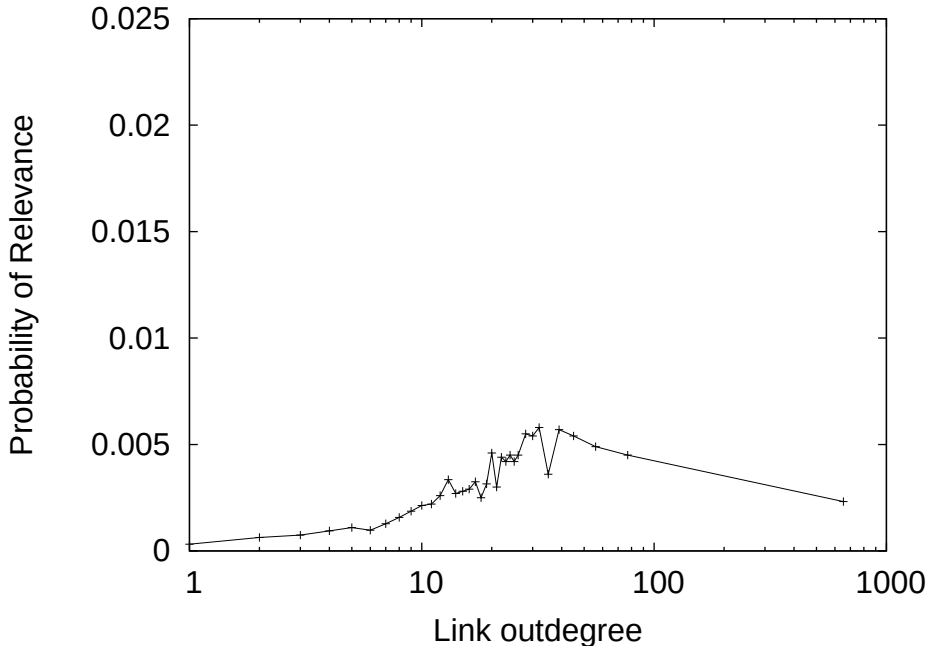
- How can we compute the probability of relevance (PoR) over link degrees?
- Recall, we use IR test-collections with topics and judgements:
  - ⋆ sort all documents on indegree (outdegree)
  - ⋆ bin per 10,000 documents
  - ⋆ count documents in each bin relevant for one of the topics
  - ⋆ PoR is the ratio of relevant documents in each bin

Computing Probability of Relevance

- How can we compute the probability of relevance (PoR) over link degrees?
- Recall, we use IR test-collections with topics and judgements:
  - ⋆ sort all documents on indegree (outdegree)
  - ⋆ bin per 10,000 documents
  - ⋆ count documents in each bin relevant for one of the topics
  - ⋆ PoR is the ratio of relevant documents in each bin
- If the link degree (in- or out-) is related to relevance, we expect to see the PoR go either up or down with increasing degree

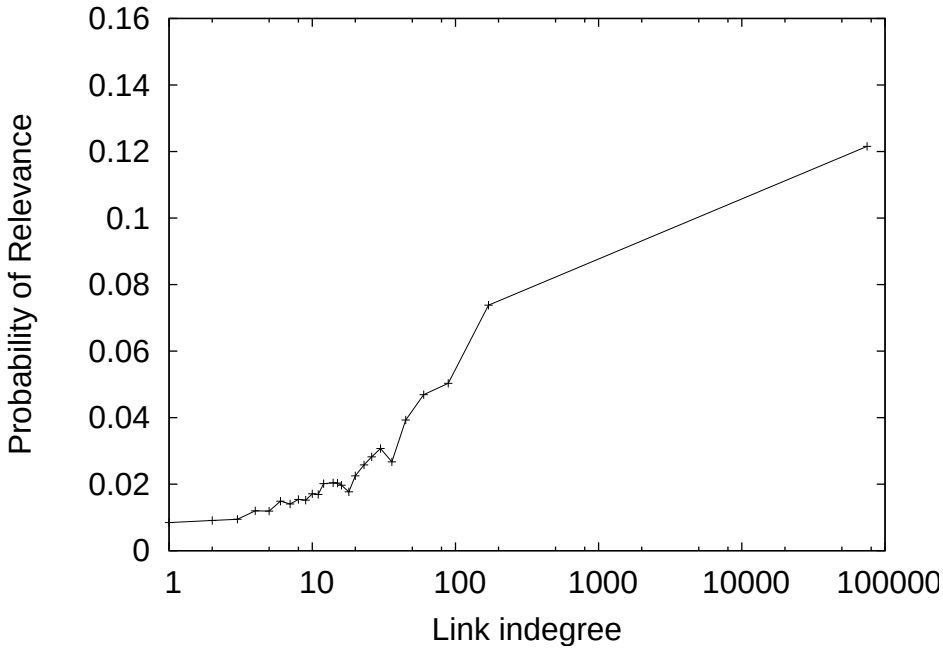
Probability of Relevance in .GOV

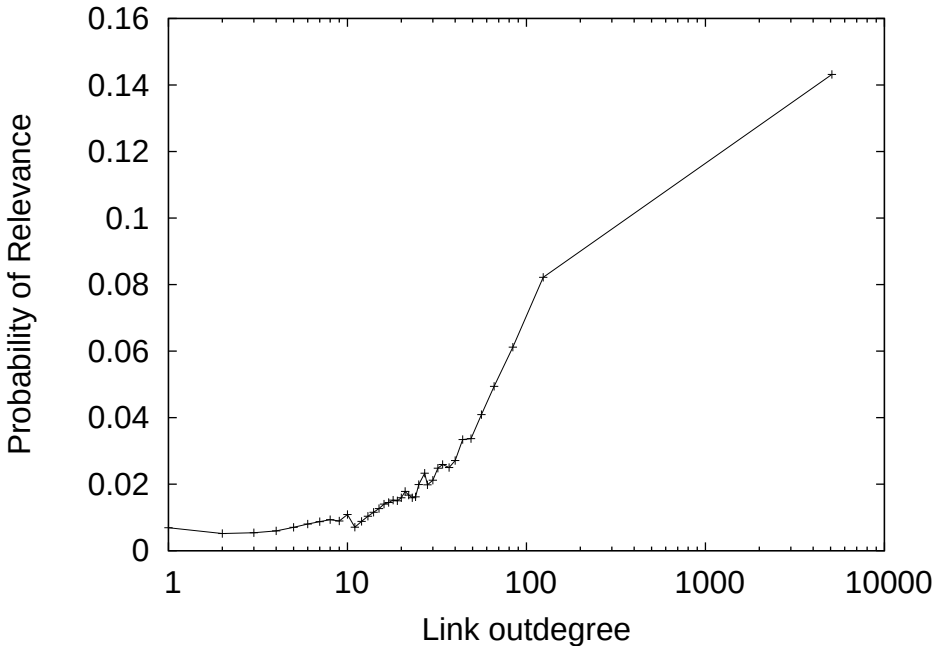




- Plots show probability of relevance over Indegree (left) and outdegree (right)
  - ⋆ Prob. of rel. goes up with increasing indegree
  - ⋆ Prob. of rel. relation with outdegree is much less clear

Probability of Relevance in Wikipedia





- Plots show probability of relevance over Indegree (left) and outdegree (right)
  - ⋆ Prob. of rel. goes up with increasing indegree and outdegree

Wrap up: Web and Wikipedia Link Topology

- How does the link topology relate to the relevance of retrieval results?
  - ⋆ for Web, indegree clearly has a relation to relevance
  - ⋆ outdegree has much weaker relation to relevance

Wrap up: Web and Wikipedia Link Topology

- How does the link topology relate to the relevance of retrieval results?
  - ⋆ for Web, indegree clearly has a relation to relevance
  - ⋆ outdegree has much weaker relation to relevance
  - ⋆ for Wikipedia, indegree and outdegree show a similar, clear relation to relevance

Outline

- Introduction
- Comparative Analysis of Link Structure
  - ⋆ Are there differences between degree distributions of incoming and outgoing links?
  - ⋆ How does the link topology relate to relevance of retrieval results?
- Link Evidence in Retrieval
  - ⋆ What is the impact of link evidence on effectiveness of Wikipedia and Web retrieval?
- Conclusions

Using Only Link Evidence in Wikipedia

Title	Global indegree	Title	Local indegree
Test cricket	1405	Toy Story	33
Nobel Prize in Physics	557	Toy Story 2	22
Sequel	529	Pixar	20
1999 in film	427	Buzz Lightyear	8
Jet Engine	341	Cars (film)	6

- Example topic *Toy Story*
  - ⋆ Global degree of top results leads to infiltration of off-topic pages with high global degrees

Using Only Link Evidence in Wikipedia

Title	Global indegree	Title	Local indegree
Test cricket	1405	Toy Story	33
Nobel Prize in Physics	557	Toy Story 2	22
Sequel	529	Pixar	20
1999 in film	427	Buzz Lightyear	8
Jet Engine	341	Cars (film)	6

- Example topic *Toy Story*
  - ⋆ Global degree of top results leads to infiltration of off-topic pages with high global degrees
- Therefore, we also look at local link evidence:
  - ⋆ links in top 100 results for the query, thus query dependent!
  - ⋆ keeps better focus on the topic of request
  - ⋆ careful how use link evidence!

Using Only Link Evidence in .GOV

Title	Global indegree	Title	Local indegree
Site Map	3,119	Bureau of Labor Statistics Home Page	61
Online Library - HUD	2,119	NTP Meetings & Events	58
Bureau of Labor Statistics Home Page	1,119	Recalls and other Press Releases	5
AMS - Search	730	What's New	3
The United States Mint	722	NCDC: Climate of 2001 - Climate Perspectives Reports	3

- Example topic *Groundhog Day Punxsutawney*
  - ⋆ Global link evidence leads to infiltration of unrelated pages with high global degrees
  - ⋆ Local link evidence leads to infiltration as well, but some loosely related pages

Using Link Evidence in a Retrieval Model

- We use standard language models:

$$P(d|q) = P(d) \cdot \prod_{t \in q} ((1 - \lambda) \cdot P(t|D) + \lambda \cdot P(t|d))$$

where $P(d)$ is a document prior to incorporate link evidence

Using Link Evidence in a Retrieval Model

- We use standard language models:

$$P(d|q) = P(d) \cdot \prod_{t \in q} ((1 - \lambda) \cdot P(t|D) + \lambda \cdot P(t|d))$$

  where $P(d)$ is a document prior to incorporate link evidence

- Standard prior

$$P(d) \propto 1 + \text{LinkDegree}(d)$$

Using Link Evidence in a Retrieval Model

- We use standard language models:

$$P(d|q) = P(d) \cdot \prod_{t \in q} ((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d))$$

  where $P(d)$ is a document prior to incorporate link evidence

- Standard prior

$$P(d) \propto 1 + \text{LinkDegree}(d)$$

- More careful Logged prior

$$P_{\log}(d) \propto 1 + log(1 + \text{LinkDegree}(d))$$

Global Link Priors

Run id	.GOV MAP	.GOV % change	Wiki MAP	Wiki % change
baseline	0.3970	–	0.3090	–

Global Link Priors

Run id	.GOV MAP	.GOV % change	Wiki MAP	Wiki % change
baseline	0.3970	–	0.3090	–
indegree	**0.4738***	+ 19.35	0.3018^{-}	- 2.33
outdegree	0.4299*	+ 8.29	0.3016^{-}	- 2.39
log(indegree)	0.4449*	+ 12.07	0.2865^{-}	- 7.28
log(outdegree)	0.4082*	+ 2.82	0.2890^{-}	- 6.47

- We see:
  - ⋆ For Web, outdegree is effective
  - ⋆ Indegree is even more effective
  - ⋆ Log prior less effective (no need to be careful)

Global Link Priors

Run id	.GOV MAP	.GOV % change	Wiki MAP	Wiki % change
baseline	0.3970	–	0.3090	–
indegree	**0.4738***	+ 19.35	0.3018^{-}	- 2.33
outdegree	0.4299*	+ 8.29	0.3016^{-}	- 2.39
log(indegree)	0.4449*	+ 12.07	0.2865^{-}	- 7.28
log(outdegree)	0.4082*	+ 2.82	0.2890^{-}	- 6.47

- We see:
  - ⋆ For Web, outdegree is effective
  - ⋆ Indegree is even more effective
  - ⋆ Log prior less effective (no need to be careful)
  - ⋆ For Wikipedia, global link evidence hurts performance
  - ⋆ Outdegree same effect as indegree

Local Link Priors

Run id	.GOV		Wiki	
	MAP	% change	MAP	% change
baseline	0.3970	–	0.3090	–

Local Link Priors

Run id	.GOV MAP	.GOV % change	Wiki MAP	Wiki % change
baseline	0.3970	–	0.3090	–
indegree	**0.4799**⋆	+ 20.88	0.3190⋆	+ 3.24
outdegree	0.4497⋆	+ 13.27	**0.3199**⋆	+ 3.53
log(indegree)	0.4410⋆	+ 11.08	0.3176⋆	+ 2.78
log(outdegree)	0.4181⋆	+ 5.31	0.3156⋆	+ 2.14

- We see:
  - ⋆ For Web, local has similar effect as global link evidence:
  - ⋆ Outdegree effective, Indegree more effective
  - ⋆ Log prior less effective (no need to be careful)

Local Link Priors

Run id	.GOV MAP	.GOV % change	Wiki MAP	Wiki % change
baseline	0.3970	–	0.3090	–
indegree	**0.4799**⋆	+ 20.88	0.3190⋆	+ 3.24
outdegree	0.4497⋆	+ 13.27	**0.3199**⋆	+ 3.53
log(indegree)	0.4410⋆	+ 11.08	0.3176⋆	+ 2.78
log(outdegree)	0.4181⋆	+ 5.31	0.3156⋆	+ 2.14

- We see:
  - ⋆ For Web, local has similar effect as global link evidence:
  - ⋆ Outdegree effective, Indegree more effective
  - ⋆ Log prior less effective (no need to be careful)
  - ⋆ For Wikipedia, local link evidence does improve performance
  - ⋆ Again, outdegree same effect as indegree

Conclusions

- Is Wikipedia link structure different?
  - ⋆ Comparative analysis of link structure:
    - ∗ Wikipedia is more densely linked than .GOV
    - ∗ for both collections, indegree is indicator of relevance
    - ∗ for Wikipedia, outdegree is also related relevance
  - ⋆ Link evidence in retrieval:
    - ∗ For Web, indegree more effective than outdegree, global degree similar to local degree
    - ∗ For Wikipedia, indegree and outdegree equally effective, but only when taking local context into account

Thank You!

References

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th International World-Wide Web conference WWW9*, pages 309–320. Elsevier Science, Amsterdam, 2000.

D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005.

J. Kamps and M. Koolen. The importance of link evidence in Wikipedia. In *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008)*, volume 4956 of *Lecture Notes in Computer Science*, pages 270–282. Springer Verlag, Heidelberg, 2008.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999. ISSN 0004-5411. doi: http://doi.acm.org/10.1145/324133.324140.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. URL citeseer.ist.psu.edu/page98pagerank.html.

I. Soboroff. Do trec web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792554.