Link-based and Content-based Evidential Information in a Belief Network Model

I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, N. Ziviani

Best Student Paper in SIGIR '2000

Ruey-Lung, Hsiao

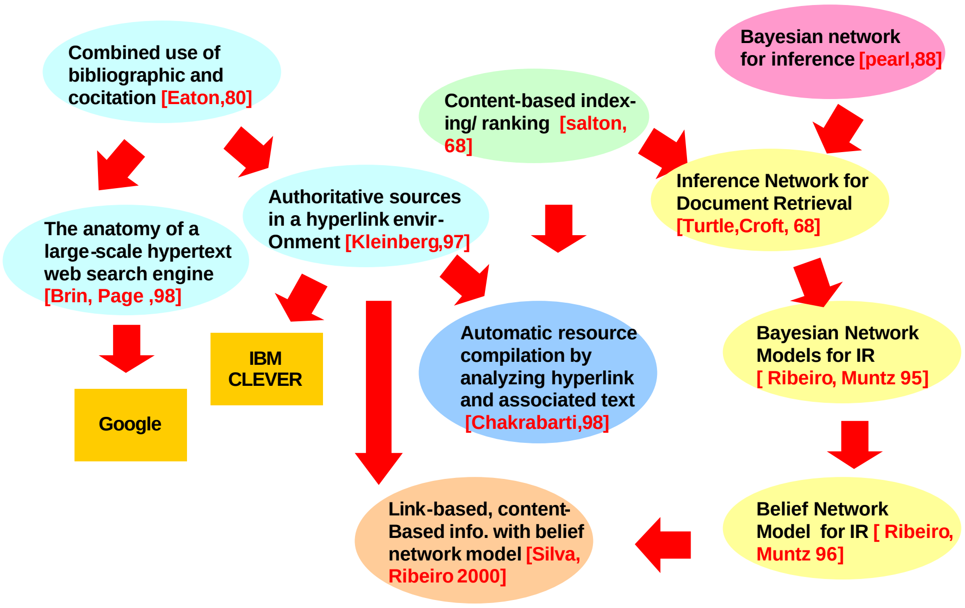
presented on Oct 11 , 2000

Introduction

- **Strategies to determine the ranking of documents in Web Search Engine**
  - Content-Based
  - Link-based
  - Combination of Content-based and Link-based
- **Inference Network / Belief Network Model**
  - Can be used as a general framework for classical IR
  - Allows combining features of distinct models into the same representation scheme

In this paper, the authors purpose a retrieval model, which provides a framework for combining information extracted from the content of the documents with information derived from cross-references among the documents, based on belief network model.

History



Combined use of bibliographic and cocitation [Eaton,80]

Bayesian network for inference [pearl,88]

Content-based indexing/ ranking [salton, 68]

Authoritative sources in a hyperlink envir-Onment [Kleinberg,97]

Inference Network for Document Retrieval [Turtle,Croft, 68]

The anatomy of a large-scale hypertext web search engine [Brin, Page ,98]

IBM CLEVER

Google

Automatic resource compilation by analyzing hyperlink and associated text [Chakrabarti,98]

Bayesian Network Models for IR [Ribeiro, Muntz 95]

Link-based, content-Based info. with belief network model [Silva, Ribeiro 2000]

Belief Network Model for IR [Ribeiro, Muntz 96]

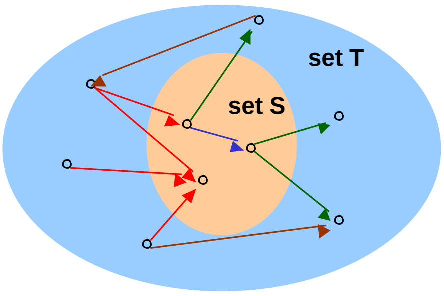
Related Work (1/4)

- Link-based information
  - Kleinberg(HITS) algorithm [kleinberg '97] [12]
    - hub/authority value for local set
  - PageRank algorithm [Brin,Page '98] [4]
- Bayesian Network Model for Information Retrieval
  - Judea Pearl purpose bayesian network to represent and infer in intelligent system. [13]
  - Turtle, Croft first use bayesian network to model information retrieval problem [19]
  - B. Ribeiro and Muntz generalize bayesian network model to be belief network model. [14,15]
- Combination of link-based/content-based information
  - Automatic resource compilation by analyzing hyperlink structure and associated text , [Chakrabarti 98] [5]
  - Improved algorithm for topic distillation in a hyperlinked environment [Bharat] [2]

Related Work (2/4)

– HITS algorithm

- Start with a root set S
  - – S_s is relatively small (typically up to 200 pages)
  - – S_s is rich in relevant pages
  - – S_s contains most (or many) of the strongest authorities.
- Recursively compute the degree of authority and hub for each element.



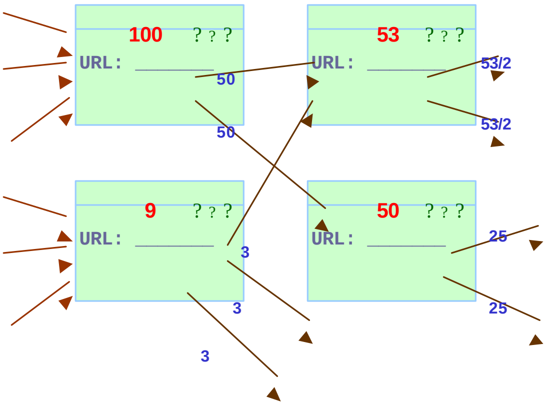
$$a(p) = ?_{q?\ p}\ h(q)$$

$$h(p) = ?_{p?\ q}\ a(q)$$

Related Work (3/4)

– PageRank algorithm

- Propagation of ranking through links

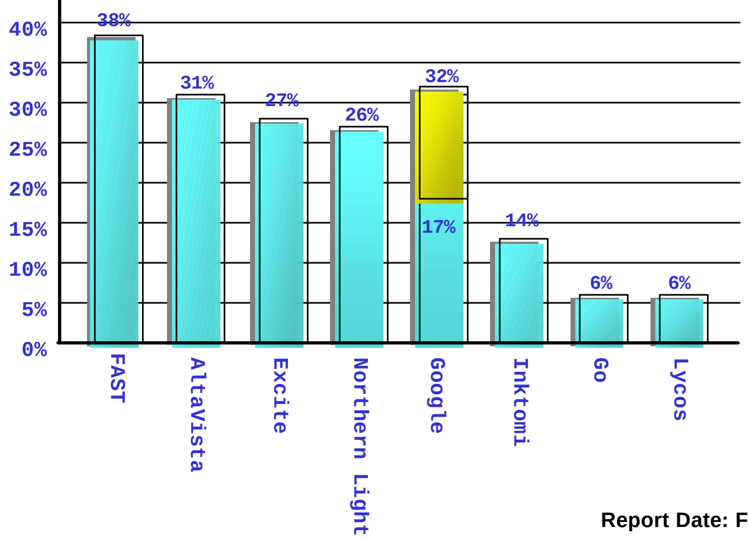


B_u : back link
F_u : forward link
N_u = | F_u |

$$R'(u) = c\ ?_{v?\ B_u} \frac{R'(v)}{N_v} + cE(u)$$

Coverage of the Web (1/2)

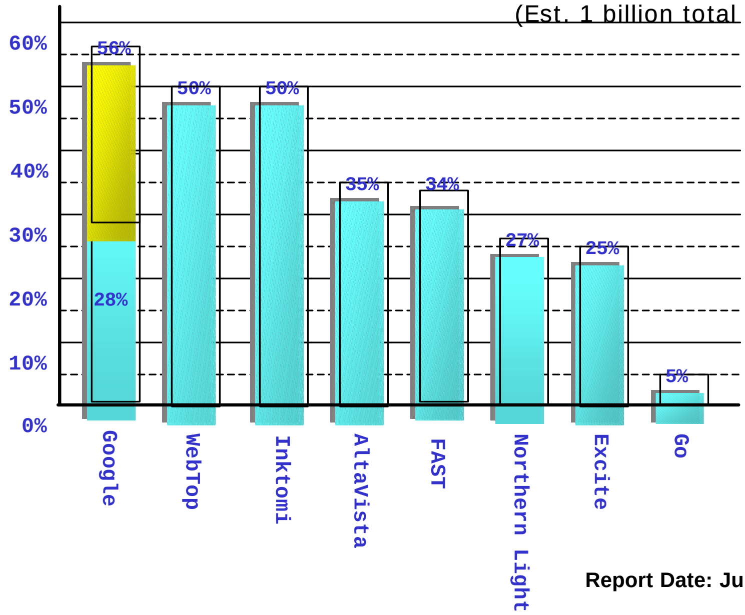
(Est. 1 billion total pages)



Report Date: Feb.3,2000

Coverage of the Web (2/2)

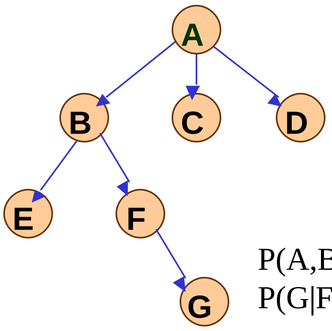
(Est. 1 billion total pages)



Report Date: Jun 6, 2000

Related Work (4/4)

- Belief Network Model
  - – Based on Bayesian Network
  - – Subsumes the classical models in IR
  - – More general than the inference network model



$X = X_1,\dots,X_n$

$P(X) = \,?_{i=1}^{n}\, P(X_i|\text{Parents}(X_i))$

P(A,B,C,D,E,F,G)=
P(G|F)P(F|B)P(E|B)P(B|A)P(C|A)P(D|A)P(A)

Belief Network Model - Ranking

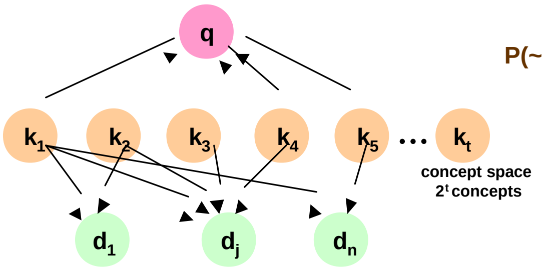
Degree of coverage of the space U by c

$P(c) = \ ?_u P(c|u) \times P(u)$

$P(u) = (\frac{1}{2})^t$

Ranking

$P(d_i|q) \ ? \ \ ?_u P(d_i|u) \times P(q|u) \times P(u)$



Vector Space Model

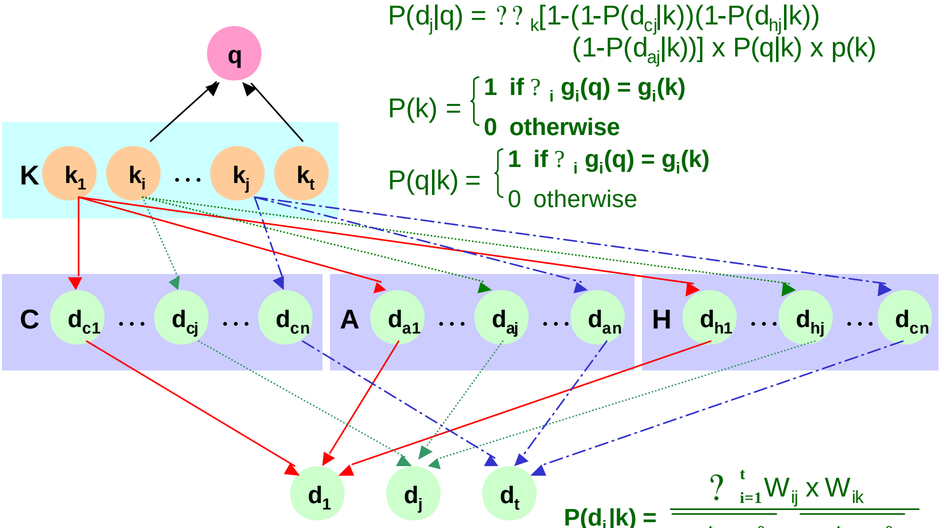
$$P(q|u) = \begin{cases} 1 \ \text{if} \ ? \ k_i, g_t(q)=g_t(u) \\ 0 \ \text{otherwise} \end{cases}$$

$P(\sim q|u) = 1 - p(q|u)$

$$P(d|u) = \frac{?\,_{i=1}^{t} W_{ij} \times W_{ik}}{?\ ?\,_{i=1}^{t} w_{ij}^{2} \ \ ?\ ?\,_{i=1}^{t} w_{ik}^{2}}$$

$P(\sim d|u) = 1 - p(d|u)$

Modeling Content/Link-Based Evidence



$$P(d_j|q) = \text{? ?}_k[1-(1-P(d_{cj}|k))(1-P(d_{hj}|k))(1-P(d_{aj}|k))] \text{ x } P(q|k) \text{ x } p(k)$$

$$P(k) = \begin{cases} \mathbf{1 \ \ if \ ?_i \ g_i(q) = g_i(k)} \\ \mathbf{0 \ \ otherwise} \end{cases}$$

$$P(q|k) = \begin{cases} \mathbf{1 \ \ if \ ?_i \ g_i(q) = g_i(k)} \\ 0 \ \ \text{otherwise} \end{cases}$$

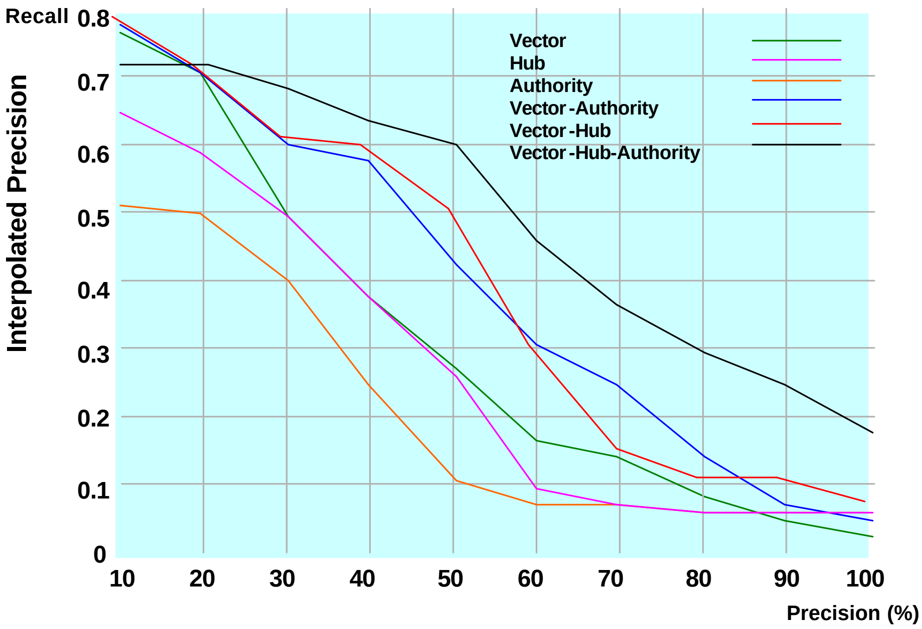
$$\mathbf{P(d_j|k)} = \frac{\text{?}\ _{i=1}^{t} W_{ij} \text{ x } W_{ik}}{\text{?}\ \overline{\text{?}\ _{i=1}^{t} W_{ij}^2}\ \text{?}\ \overline{\text{?}\ _{i=1}^{t} W_{ik}^2}}$$

Evaluation

- **Reference collection**
  - 3,027,540 pages of the Brazilian Web. (collected by CoBWeb, indexed by inverted lists)
  - 20 queries are selected from hot queries of TodoBR search engine logs.
  - For each of the 20 queries, use top 10 documents to compose query pool (so each query contains at most 60 distinct pages).
    - Average number of pages per query pool is 38.15
    - Average number of relevant pages per query pool is 17.05

Number of pages	Number of keywords	Average # of word / page	# of queries	Average # of word / query	Ave. # of page / query pool	Ave. # of relevant page / query pool
3,027,540	**3,456,910**	**512**	**20**	**1.6**	**38.15**	**17.05**

Recall ? Average precision for 20 Web queries



Conclusion

- Belief network model provides powerful mechanisms to model the information retrieval problem, specially when distinct sources of evidence are available.
- Hub and authority values performs better in combination than in isolation.

Average Precision and Gains

Recall	Vector	Vector-authority	Gain	Vector-authority	Gain	Vector-hub authority	Gain
10%	0.765	0.780	+1%	0.776	+1%	0.722	-5%
20%	0.700	0.700	+0%	0.690	-1%	0.726	+3%
30%	0.502	0.604	+20%	0.605	+20%	0.685	+36%
40%	0.366	0.574	+56%	0.591	+61%	0.640	+74%
50%	0.275	0.447	+62%	0.503	+82%	0.604	+119%
60%	0.166	0.312	+87%	0.295	+77%	0.439	+164%
70%	0.154	0.250	+62%	0.144	-6%	0.368	+138%
80%	0.080	0.144	+79%	0.098	+22%	0.297	+271%
90%	0.035	0.062	+77%	0.096	+174%	0.247	+605%
100%	0.020	0.040	+100%	0.037	+84%	0.162	+710%
Average	0.306	0.391	+27%	0.384	+25%	0.489	+59%

Reference

	Title	Author	From
Model	13. Probabilistic Reasoning in Intelligent Systems	Judea Pearl	Book 1988
	14. Bayseian network model for ir	B. Ribeiro , I. Silva	Soft Computing
	15. A belief network model for ir	B. Ribeiro , R. Muntz.	SIGIR ‘96
	19. Evaluation of an inference network-based retrieval model	H. Turtle , W. Croft	ACM trns. IS ‘91
	21. A probabilistic inference model for information retrieval.	S. Wong and Y. Yao	Info. System ‘91
Link	04. The anatomy of a large-scale hypertext web search engine	S. Brin , L. Page	WWW ‘98
	12. Authoritative sources in a hyperlinked environment.	J. M. Kleinberg	ACM-SIAM ‘98
Content	01. Modern Information Retrieval	R. Baesz-Yates, B. Ribeiro	Book ‘99
	16. Introduction to Modern Information Retrieval	G. Salton , M. McGill	Book 1983
	17. Automatic Information Organization and Retrieval	G. Salton	Book 1968
Hybrid	02. Improved algorithms for topic distillation in a hyperlink environment	K. Bharat , M. R. Henzinger	SIGIR ‘98
	05. Automatic resource compilation by analyzing hyperlink structure and associated text	G. Salton	Book 1998