Bornholm Cases in Web Mining

Finn Årup Nielsen

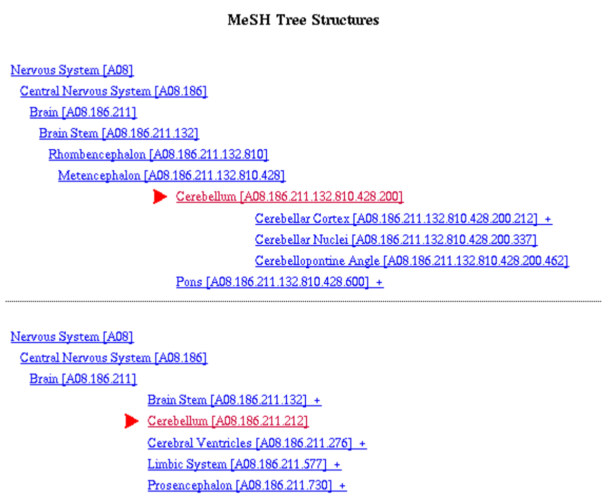
Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Email: fn@imm.dtu.dk
WWW: http://www.imm.dtu.dk/~fn

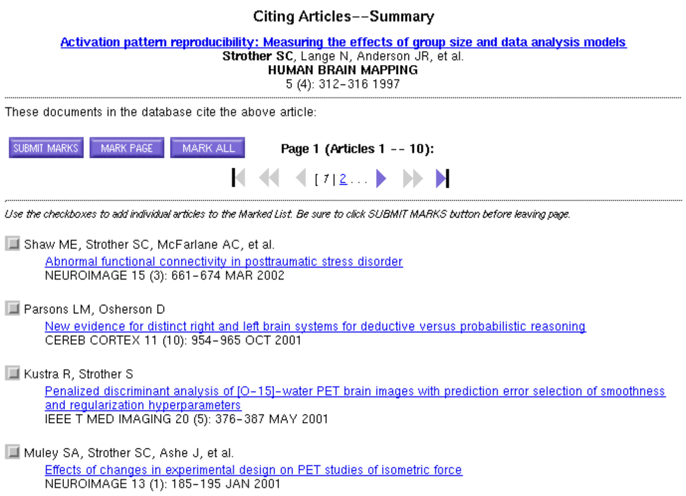
OVERVIEW

- PubMed & ISI
- ResearchIndex. Web-site with scientific computer science papers
  - Persistence of URLs.
- Inquirus, meta-search engine.
- DEADLINER. Special topic search engine.
- Webgames.
- Modeling the growth of the web.
- Gene mapping

PUBMED AND ISI



(a) PubMed



(b) ISI Web of Sience

- PubMed, http://www.ncbi.nlm.nih.gov/entrez/
  - Web-based index with bibliographic information, links to full-text. No citations.
  - Keywords in MeSH hierarchy.
  - Typed queries possible, e.g., `"Hum Brain Mapp"[Jour] AND "1997"[dp] AND "5"[vol]`.
  - "Related articles" service (Wilbur and Yang, 1996).
- ISI Web of Science, http://wos.isiglobalnet.com/
  - Web-based citation index.
  - Widely used in citation studies.

RESEARCHINDEX/CITESEER

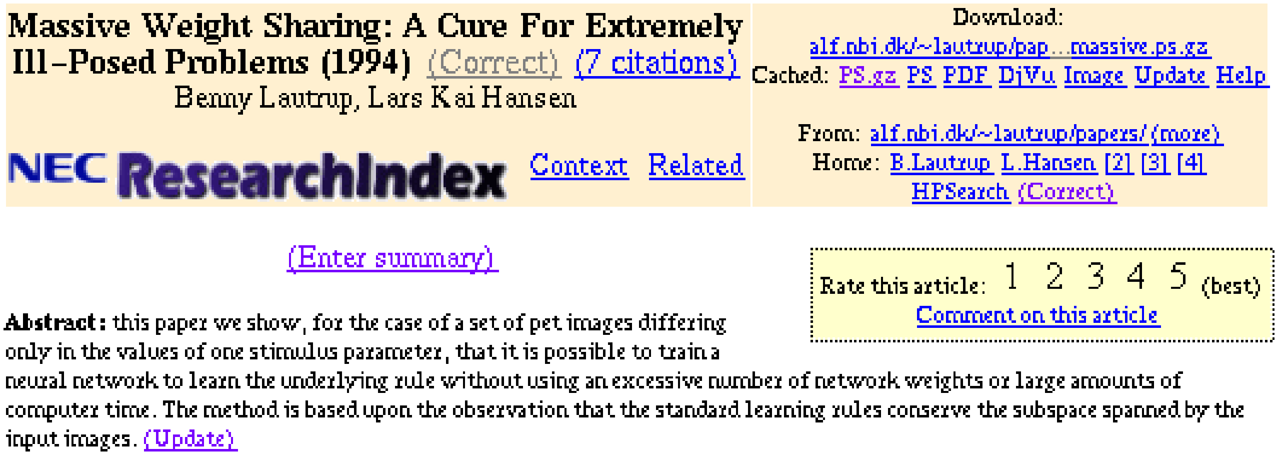


Figure 1: Screen shot from ResearchIndex.

- ResearchIndex, Citeseer. (Lawrence et al., 1999; Giles et al., 1998; Bollacker et al., 1998) http://researchindex.com.
- Collection of scientific computer science papers in PostScript/PDF files downloaded from the Internet: 270,000 articles with 100,000 internally cited.
- Conversion of PDF and PostScript files to text (and other formats)
- Identification of words and citations.
- Ad hoc text search.
- Collaborative correction (if the machine translation fails) and annotation.

RESEARCHINDEX

Cited by: More
Comparison Of Pca And Ica Based Clutter Reduction.. - Karlsen, Larsen.. (Correct)
Antenna Characteristics and Air-Ground Interface.. - Karlsen, Larsen.. (2000) (Correct)
On Independent Component Analysis for Multimedia Signals - Hansen, Larsen, Kolenda (2000) (Correct)

Similar documents (at the sentence level):
19.4%: Extremely Ill-posed Learning - Hansen, Lautrup, Law, Mørch, Thomsen (1994) (Correct)

Active bibliography (related documents): More All
0.3: Generalizable Patterns in Neuroimaging: How Many.. - Hansen, Larsen.. (1999) (Correct)
0.3: Unsupervised Learning and Generalization - Hansen, Larsen (1996) (Correct)
0.3: Nonlinear Autoregressive Moving Average Signal Modeling using the .. - Fischer (Correct)

Users who viewed this document also viewed: More All
0.0: Fast Planning Through Planning Graph Analysis - Blum, Furst (1995) (Correct)
0.0: Handwritten Digit Recognition with a Back-Propagation Network - Le Cun (1990) (Correct)
0.0: Uniform Design Based On Latin Squares - Kai-Tai Fang (Correct)

Similar documents based on text: More All
0.3: Nonlinear versus Linear Models in Functional.. - Mørch, Hansen.. (1997) (Correct)
0.3: Visualization of Neural Networks Using Saliency Maps - Mørch, Kjems, Hansen.. (1995) (Correct)
0.3: Neural Networks for Signal Processing - Svarer (Correct)

Related documents from co-citation: More All
2: Automatic Mine Detection based on Ground Penetratinig Radar (context) - Yu, Witten - 1999
2: A multivariate analysis of PET activation studies (context) - Friston, Poline et al. - 1996
2: Design and Evaluation of Neural Classifiers - Application to Skin Lesion Classif.. (context) - Hintz-Madsen, Hansen et al. - 1995

Figure 2: Screen shot from ResearchIndex.

- Sorted citation and reference list.
- Sentence hash: Identification of copied documents.
- Similar documents:
  - Text: “Similar documents based on text”.
  - Citation structure
  - Usage analysis “Users who viewed this document also viewed”.
- Similar effort
  - Cora, (McCallum et al., 2000).
  - (Bergmark, 2000): Discusses conversion to text, extraction of references and anchors, parsing of reference lists.

PERSISTENCE OF URLS

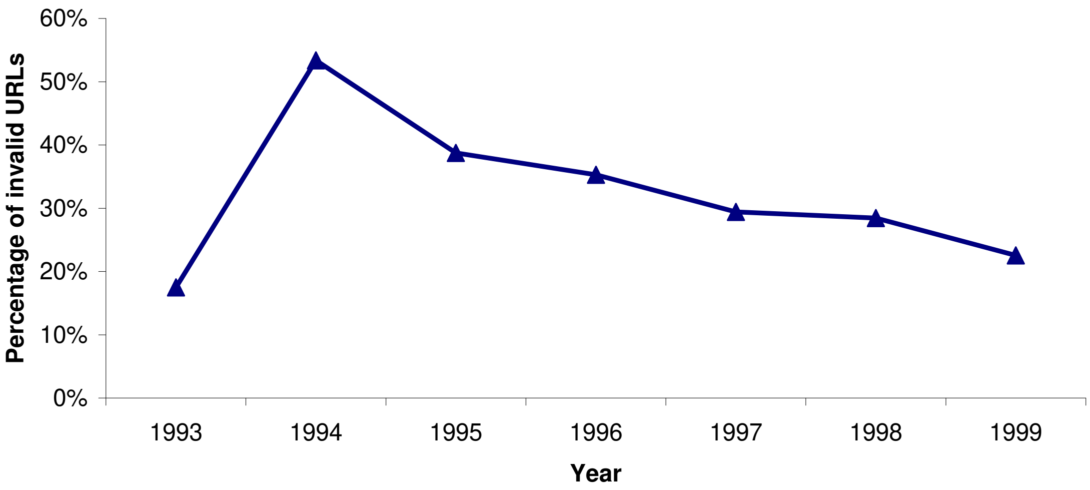


Figure 3: Percentage of invalid links (Lawrence et al., 2001, fig. 2).

- Are URLs is references in scientific documents still valid? (Lawrence et al., 2000; Lawrence et al., 2001).
- Data from ResearchIndex: 67,000 URLs extracted from 100,000 documents: Increase in average URLs from 0 in 1993 to 1.6 in 1999.
- Automatic download and if failed: Manual search for 300 randomly selected missing document by 9 people.
- Many invalid URLs (53% in 1994), personal home-page moved, servers closed.
- Often easy to relocate with search engine or guessing the URL and sometimes the URL is not important.

INQUIRUS & INQUIRUS 2

- Inquirus meta-search engine (Lawrence and Giles, 1998).
  - Aggregation of results from several search engines.
  - Identification of hubs and authorities.
- "Inquirus 2" (Glover et al., 1999; Glover et al., 2001)
  - Query modification for finding category specific web-pages, e.g., personal homepages
  - Train a classifier on a desired category (manually labeled).
  - Learn *query modificiations*, — might be different for each individual search engine
  - Personal home query modification examples: "`my "home page"`" (FAST), "`+ "my name is"`" (AltaVista), "`"s home"`" (Google).
  - A version returned results on a web-pages with Java continuously updating the ranked list.
  - An other service for query refinement (for scientific web-pages) http://www.scirus.com/

DEADLINER

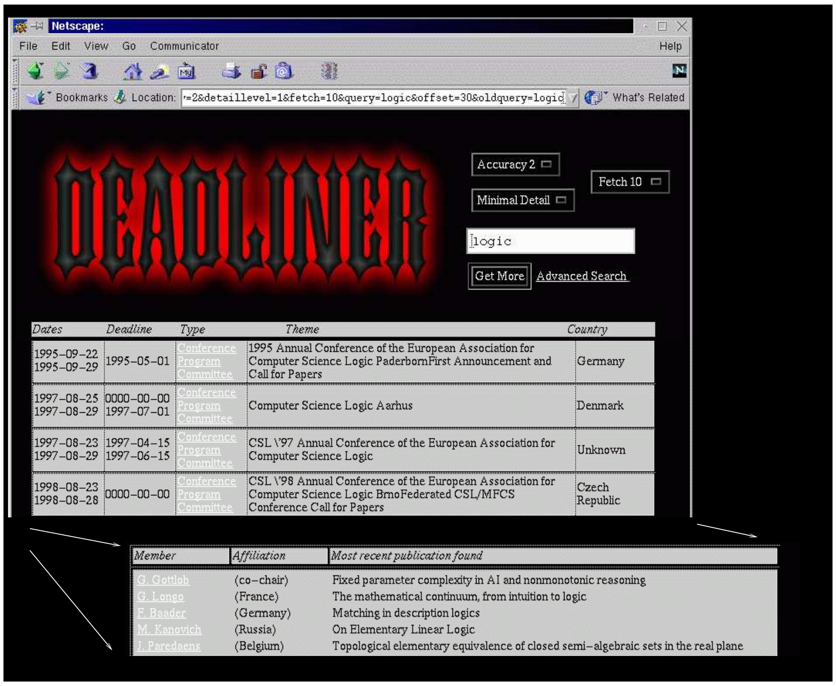


Figure 4: Screen shot from DEADLINER. From (Kruger et al., 2000, fig. 1).

- Search engine for conference and workshops (Kruger et al., 2000).
- Data from web (by query modified meta search and focused crawling), newsgroup and (emails).
- Preliminary screening with classifier.
- Superpattern matching on personal names with dictionary lookup, e.g., list of authors extracted from ResearchIndex.
- Information is automatically entered into a MySQL database.
- Other example: http://www.campsearch.com

WEBGAMES

- Webgames are good forecasters (Pennock et al., 2001b; Pennock et al., 2001a; Pennock et al., 2001c), http://artificialmarkets.com.
- Trading with *securities*, e.g., short selling will profit if the price goes down.
- Prices predict the odds of outcome, e.g., that a horse will win.
- Market-based belief aggregatation.
- Certain knowledge is aggregated.
- Bias: Favorites are underpriced and longshots are overpriced. Risk-seeking consensus.
- The maximum price does not garanties winner, other lower priced events should come out as winner according to the price.

EXAMPLE MARKETS: HSX

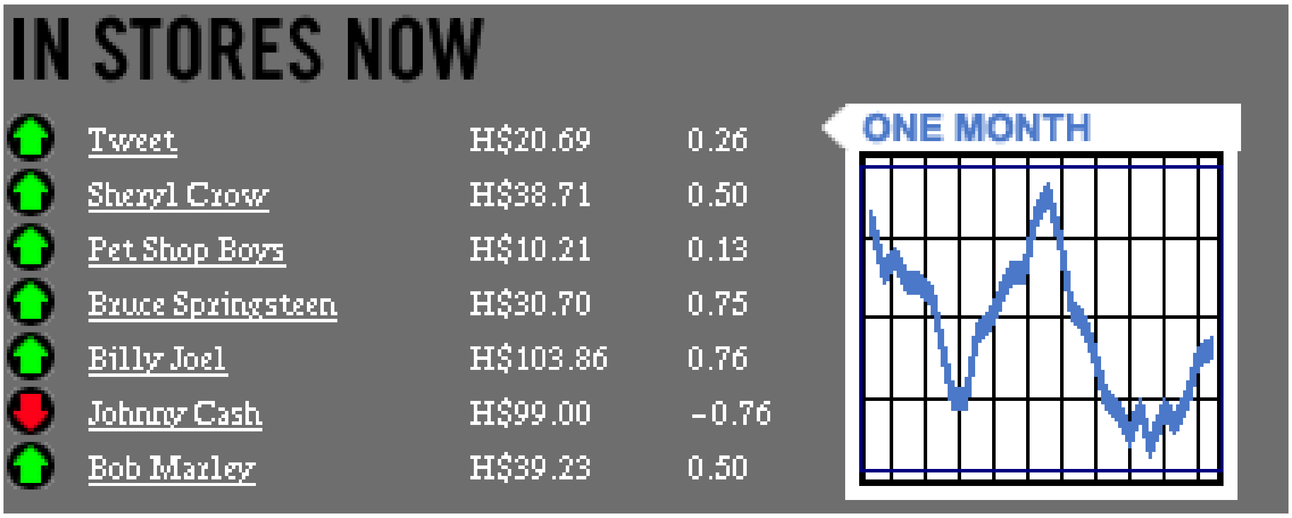


Figure 5: Screenshot from HSX with “artistocks”.

- Hollywood stock exchange (HSX) http://www.hsx.com
- Artificial money: Hollywood dollar (H$)
- Very popular with 400,000 registered accounts (in August 2000)
- Predicting box office records or award options in movies and music, e.g., MovieStock, (H$80 to $80 million in a film’s first four weeks), StarBond (Average performance of an actor or director over 5 films)
- Predictions can be more accurate than expert opinion.

EXAMPLE MARKETS: FORESIGHT EXCHANGE

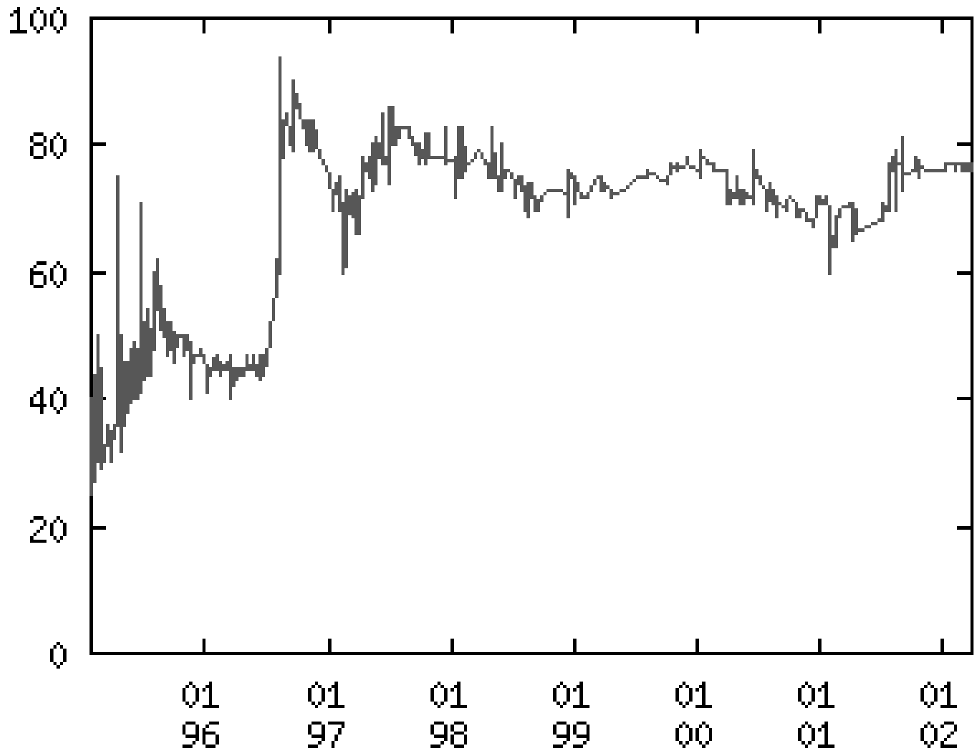


Figure 6: Price curve from Foresight Exchange for "Evidence of Extraterrestrial Life, fossils, or remains will be found by 12/31/2050". Raise in price probably caused by the "life on Mars" evidence (McKay et al., 1996).

- Foresight exchange, http://www.ideafutures.com/.
- Proposed by (Hanson, 1998; Hanson, 1995): Idea Futures market.
- Prediction in, e.g., arts, finance, news, politic, science & technology. Participants can propose.
- Examples, "Amazon crashes by 2002" (12.2%), "Human Clone before 2005" (3.6%).

EXAMPLE "MARKETS": F1P6

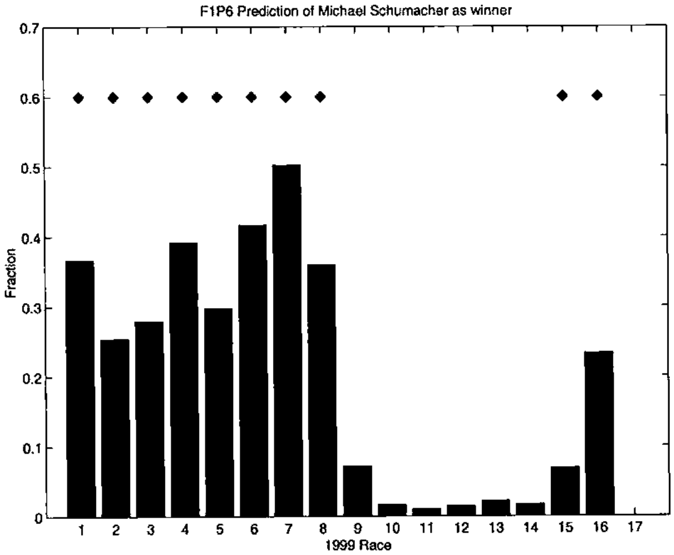


Figure 7: Not fully informed: Prediction of Michael Schumacher when he was injured and out of the game show a positive probability.

- F1P6. Formula One race car prediction game via email/web. http://www.motorsport.com (Wassermann, 1996).
- Players are to predict the position of the six first drivers, e.g., SchumacherM, SchumacherR, ...
- Players are given 10 for precise prediction, 6 if one position off, and 4 (if two off), etc. for 3, 2, 1 points. Conversion from points to odds.
- Not a real market: No price!
- Associated with real money bookmakers: F1P6 compares well with these predictions.

ASSESSING PREDICTION

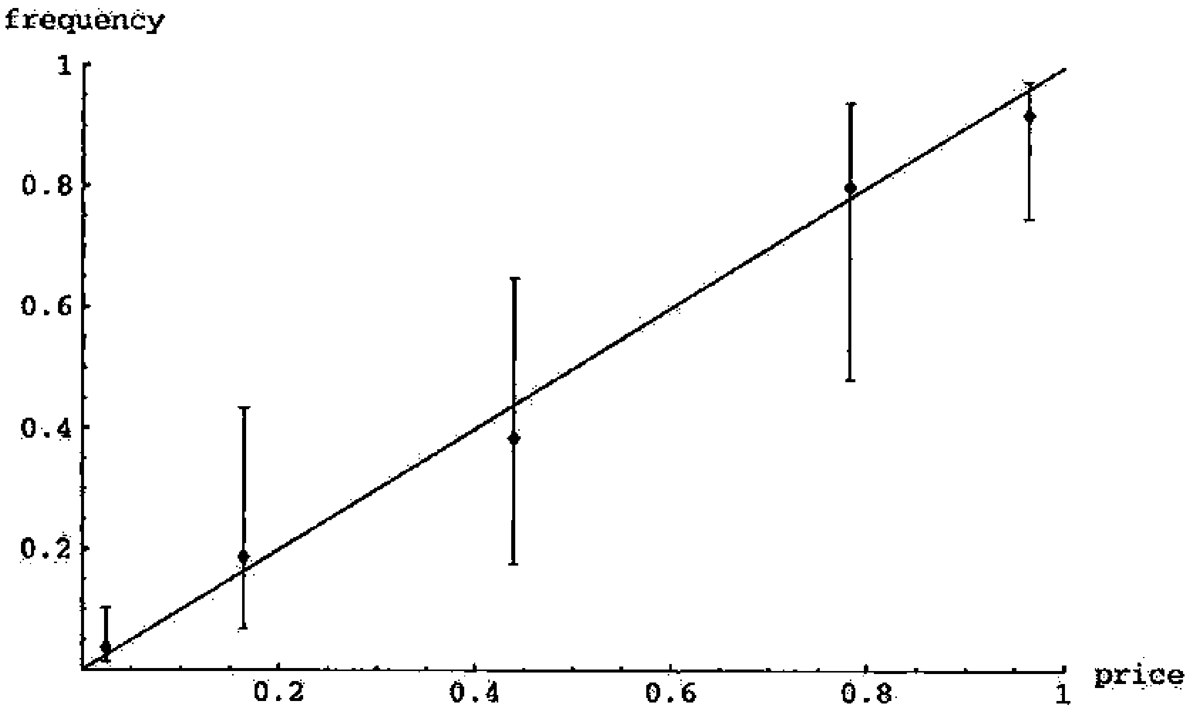


Figure 8: Accuracy of the Foresight Exchange market. Prices are 30 days before claim expiration. Points display observed frequency versus average price. The solid line corresponds to ideal accuracy. (David Pennock)

- Assessing the prediction by price with observed outcome frequency.
- Sorting by price grouping into a few number of buckets, i.e., price ranges.
- Error bars computed as regarding the outcomes as independent Bernoulli trials with a uniform prior (beta-distribution).

ASSESSING EFFICIENCY

- In an efficient market *arbitrage* should not be possible.
- Example
  - Predicting correct outcome gives 25 H$.
  - Mutually exclusive events costs 23 H$.
  - Buying stocks for all events while give certain reward.
- Unbalance in the price should by time decrease.

GROWTH OF THE WEB

- Modeling the number in-degree to a web-page (Pennock et al., 2002).
- Generative process, m_0 initial vertices, t time step adds one vertex and m edges. Probability $\Pi(k_i)$ that an edge connects to vertex i where k_i is in-degree to i

$$\Pi(k_i) = \underbrace{\alpha \frac{k_i}{2mt}}_{\text{Preferential}} + \underbrace{(1-\alpha)\frac{1}{m_0 + t}}_{\text{Uniform}} \tag{1}$$

- Fitting the model to data from Inktomi.
- From http://modelingtheweb.com

Publications	0.99	
Entertainment	0.96	More competition /
Consumer electronics	0.94	harder to enter market
Toys & games	0.92	
Antiques & collectibles	0.86	
Sports	0.76	
Health	0.75	
Home & garden	0.72	
Jewelry	0.70	
Weddings	0.65	Less competition /
Photographers	0.38	easier to enter market

PUBGENE

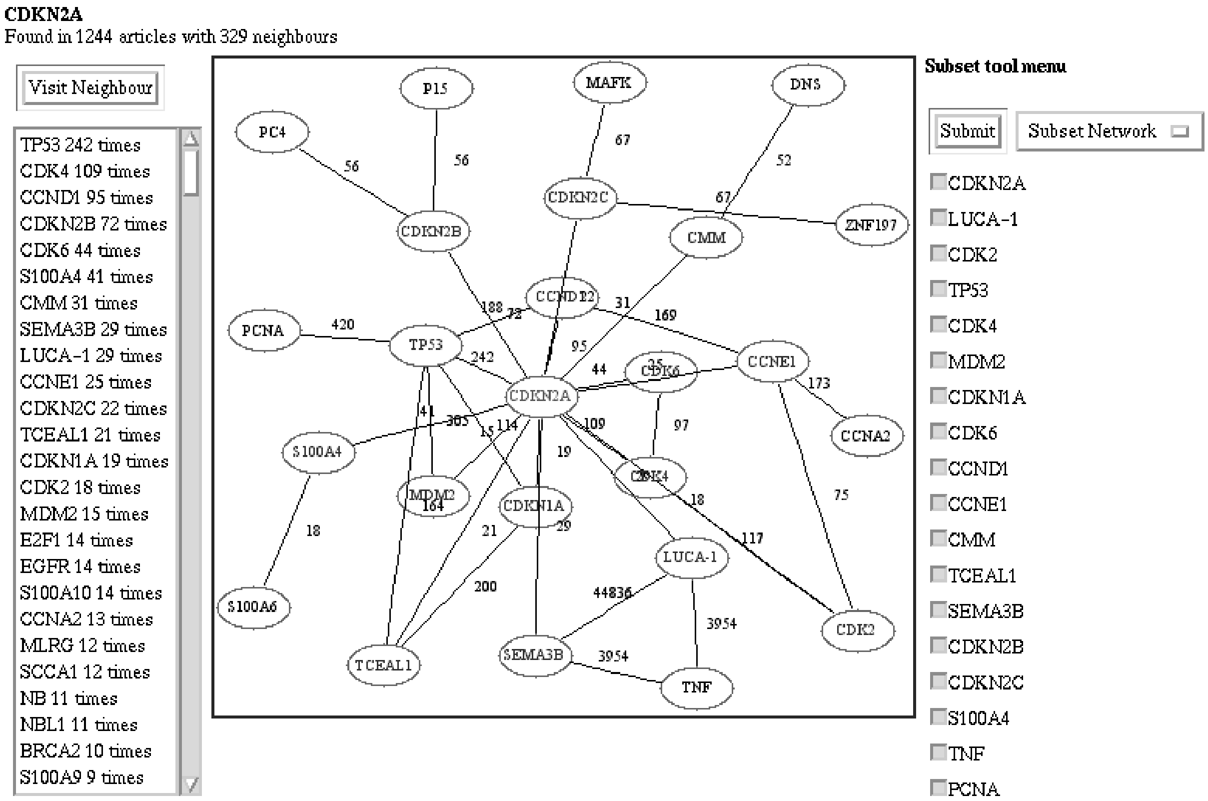


Figure 9: PubGene example for "CDK2NA".

- PubGene, http://www.pubgene.org/
- Gene cocitation analysis (Jenssen et al., 2001),
- Data from over 10 million articles from PubMed citing 13,712 human genes names obtained from HUGO, LocusLink, The Genome Database and GENATLAS.
- Network diagrams are automatically generated with the GraphViz software.

References

Bergmark, D. (2000). Automatic extraction of reference linking information from online documents. Technical Report CSTR 2000-1821, Cornell Digital Library Research Group.

Bollacker, K., Lawrence, S., and Giles, C. L. (1998). CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Agents '98, 2nd International ACM Conference on Autonomous Agents*, page 116. ACM.

Giles, C. L., Bollacker, K., and Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In Witten, I., Akscyn, R., and III, F. M. S., editors, *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA. ACM Press.

Glover, E. J., Flake, G. W., Lawrence, S., Birmingham, W. P., Kruger, A., Giles, C. L., and Pennock, D. (2001). Improving category specific web search by learning query modifications. In *Symposium on Applications and the Internet, SAINT*, pages 23–31.

Glover, E. J., Lawrence, S., Birmingham, W. P., and Giles, C. L. (1999). Architecture of a metasearch engine that supports user information needs. In *Eighth International Conference on Information and Knowledge Management (CIKM 99)*.

Hanson, R. (1998). Decision markets. *IEEE Intelligent Systems*, 14(3):16–19.

Hanson, R. D. (1995). Could gambling save science? encouraging an honest consensus. *Social Epistemology*, 9(1):3–33.

Jenssen, T.-K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28. PMID: 11326270. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v28/n1/full/ng0501_21.html. Gene cocitation analysis involving data from over 10 million articles from PubMed citing 13,712 human genes names obtained from HUGO, LocusLink, The Genome Database and GENATLAS. Network diagrams are automatically generated with the GraphViz software.

Kruger, A., Giles, C. L., Coetzee, F. M., Glover, E., Flake, G. W., Lawrence, S., and Omlin, C. (2000). DEADLINER: Building a new niche search engine. In *Ninth International Conference on Information and Knowledge Management, CIKM 2000, Washington, DC*.

Lawrence, S., Coetzee, F., Flake, G. W., Pennock, D. M., Krovetz, R., Nielsen, F. Å., Kruger, A., and Giles, C. L. (2000). Persistence of information on the web: Analyzing citations contained in research article. In *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 6-11, 2000*, pages 235–242. ACM.

Lawrence, S., Coetzee, F., Glover, E., Pennock, D. M., Flake, G., Nielsen, F. Å., Krovetz, R., Kruger, A., and Giles, C. L. (2001). Persistence of web references in scientific research. *IEEE Computer*, 34(2):26–31. http://www.neci.nec.com/~lawrence/papers/persistence-computer01/persistence-computer01.ps.gz.

Lawrence, S. and Giles, C. L. (1998). Context and page analysis for improved web search. *IEEE Internet Computing*, 2(4):38–46.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71.

McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163. http://www.cs.cmu.edu/ knigam/papers/cora-jnl.ps.gz. ResearchIndex: http://citeseer.nj.nec.com/mccallum99automating.html.

McKay, D. S., Jr., E. K. G., Thomas-Keprta, K. L., Vali, H., Romanek, C. S., Clemett, S. J., Chillier, X. D. F., Maechling, C. R., and Zare, R. N. (1996). Search for past life on mars: Possible relic biogenic activity in martian meteorite ALH84001. *Science*, 273(5277):924–30. PMID: 8688069.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211. http://modelingtheweb.com/modelingtheweb.ps.

Pennock, D. M., Lawrence, S., Giles, C. L., and Nielsen, F. Å. (2001a). Extracting collective probabilistic forecasts from web games. In *KDD-2001, The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26 - 29, 2001, San Francisco, California, USA*. SIGKDD, Association for Computing Machinery.

Pennock, D. M., Lawrence, S., Giles, C. L., and Nielsen, F. Å. (2001b). The power of play: Efficiency and forecast accuracy in web market games. Technical Report 2000-168, NEC Research Institute, 4 Independence Way, Princeton, New Jersey.

Pennock, D. M., Lawrence, S., Giles, C. L., and Nielsen, F. Å. (2001c). The real power of artificial markets. *Science*, 291(5506):987–988.

Wassermann, D. A. (1996). More than enough answers to questions nobody asked about F1P6. http://www.freenet.edmonton.ab.ca/~davidwss.

Wilbur, W. J. and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, 26(3):209–222. PMID: 8725772.