Information Genealogy: Uncovering the Flow of Ideas in Non-Hyperlinked Document Databases

Benyah Shaparenko, Thorsten Joachims
KDD 2007

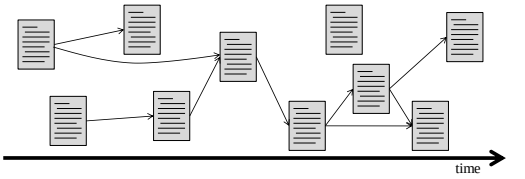
Thorsten Joachims
Cornell University
Based on slides by Benyah Shaparenko

Archives

Motivation: We now have more then >10 years of online

- Newspaper archives
- Conference proceeding
- Personal email and photos
- Blogs, Wikipedia(?), etc.

- **Archival, self-referential process of corpus development**



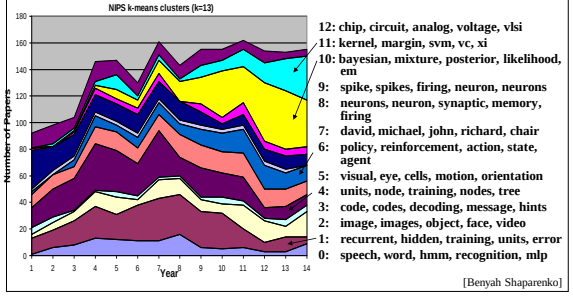
Possible Research Questions

- How did the topics in the corpus change over time?
- What are articles are related?
- Did one article influence another article?
- Who were the most influential authors?
- Who are the bloggers that are ahead of the curve?
- An automatic personal diary from email and photos.
- News: New stories identification. Remove redundancy
- Reflection: how do you spend your time/money.
- Social influence, how do stories travel.
- Photos to stories, reduce information. Your year in photos.
- Speed up desktop search, make interactive.
- Temporal representations as a way of organizing search.
- Collaborative Search, use other peoples traces.
- Time-aware search, consistency across corpora
- Self-organizing encyclopedia, multi-media
- Predicting trends, life-cycle
- What blogs are hot, personal interest.
- Visualizing social network
- Categorizing images, use the many images on the internet.
- Questions qnswering
- Handling analogy in search
- Google squared
- Evolution of information, wikipedia
- Trends and relationships between trends
- Changes of scene over time (time travel in images)
- Relative time in time (use time as part of query)
- Search as a zoom of a collection
- Why are we storing archives? Events, personalities, Change of personality

Ideas: Understanding Archives

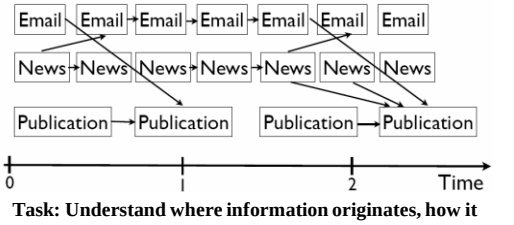
- **Visualization of content**
  - Over time / landmarks / your year in photos / zoom content
- **Summarization / aggregation of content**
  - Summary of collection / Wikipedia curation / sentiment
- **Extract temporal development of content**
  - Trends / what is hot
- **Augment collection with structure**
  - identify relationships between documents / dependencies between documents and authors, institutions, … / influence
- **Personal information management**
  - Search with support for time
  - Photo archives / diary / reflection / where do I spend my time

Summarizing Temporal Development: Neural Information Processing Systems (NIPS) 1987 - 2000



[Benyah Shaparenko]

Information Genealogy



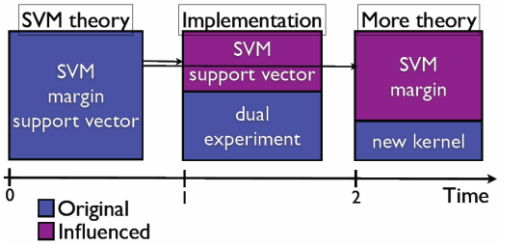
Task: Understand where information originates, how it spreads, and how information streams evolve over time.

- How did the ideas in a collection evolve?
- Who were the most influential authors driving the change?
- Did one news article influence another article?
- Who are the bloggers that are ahead of the curve?

Questions

- **How did ideas develop and spread in a given corpus?**
- **What are the inter-document influence relationships through which ideas spread?**
- **Which documents are most influential?**

Model of Influence



- **Key ideas and modeling assumptions**
  - No explicit citation/hyperlink structure
  - Influence is encoded in statistical signatures of word use
  - Topical similarity is not equal to influence

Related Work

- **Topic Detection and Tracking (e.g. Allan/et al./98)**
- **Real-world Influence on Documents (Kleinberg/02)**
- **Citation and Hyperlink Analysis (e.g. Kleinberg/99, Page/Brin/98, Garfield/03)**
- **Automatic Hypertext and Link Detection (e.g. Allan/et al/98)**
- **Language and Topic Modeling (e.g. Steyvers/et al/04, Hofmann/98, Kurland/Lee/04)**

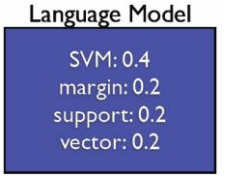
Generative Model of Corpus

Generative Modeling Assumptions:

- **Documents are generated as probabilistic mixtures of previous documents and original ideas**
- **Measure influence by how much documents base their content on previous documents**

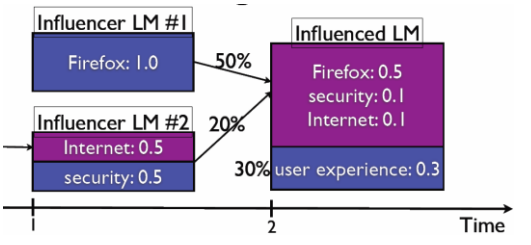
Modeling Documents

- **Unigram language model**
- **Document is a vector-valued random variable $D=(W_1,\ldots,D_l)$**
- **Generate document by drawing i.i.d. from language model θ**



$$P(D^{(i)} = d^{(i)} \mid \theta^{(i)}) = P(D^{(i)} = (w_1^{(i)} \cdots w_{|D^{(i)}|}^{(i)}) \mid \theta^{(i)})$$
$$= \prod_{j=1}^{|D^{(i)}|} P(W^{(i)} = w_j^{(i)} \mid \theta^{(i)}) = \prod_{j=1}^{|D^{(i)}|} \theta_{w_j}^{(i)}$$

Modeling Influence



- **Document language models are a mixture of the language model of its influencers, plus an original part.**

Inter-Document Influence Model

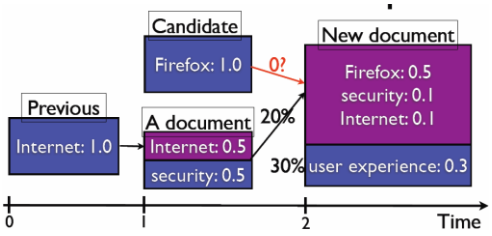
- **Influence: A document's language model is given by a mixture of preceding document's language models.**

$$P(D^{(i)} = d^{(i)} \mid \pi^{(i)}) = \prod_{j=1}^{n_i} \left(\pi_n^{(i)} \bar{\theta}_{w_j^{(i)}}^{(i)} + \sum_{k \in \mathcal{P}} \pi_k^{(i)} \hat{\theta}_{w_j^{(i)}}^{(i)} \right)$$

$$0 \leq \pi_k^{(i)}, \pi_n^{(i)} \text{ and } \pi_n^{(i)} + \sum_k \pi_k^{(i)} = 1$$

- **Note: Only temporally preceding documents can influence this document.**

Question: How can we Detect Influence?



- **Hypothesis Test**
  - Null Hypothesis: Candidate document has mixing weight 0.
  - Alt. Hypothesis: Candidate has positive mixing weight.

Likelihood Ratio Test for Influence

- **Space of all mixtures models**

$$\Pi = \left\{ \pi^{(new)} : \pi_{can}^{(new)} + \sum_{k \in \mathcal{P}} \pi_k^{(new)} = 1 \ \wedge \ \pi_k^{(new)} \geq 0 \ \wedge \ \pi_{can}^{(new)} \geq 0 \right\}$$

- **Null Hypothesis: Candidate document has no influence (i.e. mixing weight 0).**

  →Space of mixture models restricted to those consistent with null hypothesis

$$\Pi_0 = \left\{ \pi^{(new)} : \pi_{can}^{(new)} + \sum_{k \in \mathcal{P}} \pi_k^{(new)} = 1 \ \wedge \ \pi_k^{(new)} \geq 0 \ \wedge \ \pi_{can}^{(new)} = 0 \right\}$$

- **Statistic:** $\Lambda_{d^{(can)}}(d^{(new)}) = \frac{\sup_{\pi \in \Pi_0} \{P(D^{(new)} = d^{(new)} | \pi)\}}{\sup_{\pi' \in \Pi} \{P(D^{(new)} = d^{(new)} | \pi')\}}$

- **Reject null hypothesis if** $-2 \log(\Lambda_{d^{(can)}}(d^{(new)})) > c$

Computing the LRT

- **Two optimization problem per LRT**
- **Maximize likelihood L for parameters in S**
- **Optimization Problem:**

$$\max_{\pi \in \Re^{|S|}} \quad \log L(\pi \mid d^{(new)})$$

$$\text{subject to} \quad \sum_{k \in \mathcal{S}} \pi_k^{(new)} = 1$$

$$\forall k \in \mathcal{S} : \pi_k^{(new)} \geq 0$$

  →Convex (no local optima)

- **Heuristic: Consider only documents that are sufficiently similar.**

Experiments

- **Can we derive an influence graph from non-hyperlinked text?**
- **Can we identify the most influential documents?**

Identifying Dependencies and Influence

Which papers were influenced by "Shrinking the Tube: a New Support Vector Regression Algorithm" written by B. Schoelkopf et al.?

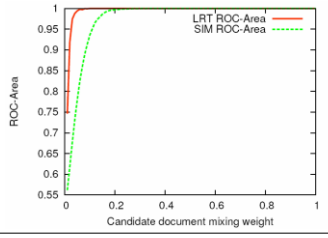
- Assume unigram word distribution is mixture of past papers
- Likelihood ratio test for non-zero mixture weight (convex program)

log(Λ(d))	Cite?	Title and Authors
321.2	No	"Support Vector Method for Novelty Detection", B. Schoelkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt.
221.8	Yes	"An Improved Decomposition Algorithm for Regression Support Vector Machines", Pavel Laskov.
219.9	Yes	"v-arc: Ensemble Learning in the Presence of Outliers", G. Raetsch, B. Schoelkopf, A. Smola, K. Miller, T. Onoda, S. Mims.
184.6	No	"Fast Training of Support Vector Classifiers", F. Perez-Cruz, P. Alarcon-Diana, A. Navia-Vazquez, A. Artes-Rodriguez.
168.9	Yes	"Uniqueness of the SVM Solution", C. Burges, D. Crisp.

[ShapaJo07]

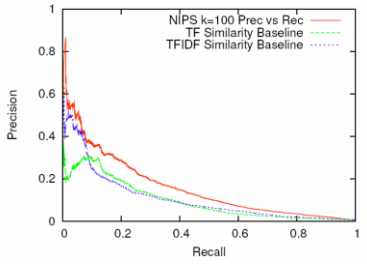
Influence Graph: How sensitive is the Test?

- **Data:**
  - Synthetic data generated according to mixture model.
  - Base language models are taken from random NIPS documents.



ROC-Area@10%	
Score	Weight
500.3	10% (+)
100.5	10% (+)
25.2	10% (+)
7.3	0% (-)
3.6	10% (+)
2.8	0% (-)
1.1	0% (-)
0.9	0% (-)

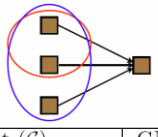
Influence Graph: Quality of the Predicted Influence Links



Precision Recall	
Score	Cite?
500.3	+
100.5	+
25.2	+
7.3	-
3.6	+
2.8	-
1.1	-
0.9	-

Impact of Similarity Heuristic

- **Experiment:**
  - Condition 1: Use pre-selection based on similarity
  - Condition 2: Make sure all cited documents are included.



Arrows = citations
Red: Candidate set
Blue: Incl. all citations

Dataset ($\mathcal{C}$)	GMAP	GMAP (perfect $\mathcal{C}$)
NIPS (TFIDF)	0.4531	0.4556
NIPS (TF)	0.4489	0.4590
HEPTH (TFIDF)	0.2543	0.3803
HEPTH (TF)	0.2432	0.3906

Key Documents by Year: NIPS

- **Influence: In-degree from Influence Graph**

Document		Citation Counts	
Year	Document Title and Author(s)	NIPS	Google Scholar
1988	"Efficient Parallel Learning Algorithms for Neural Networks" by Alan Kramer, A. Sangiovanni-Vincentelli	2	89
1989	"Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters" by John S. Bridle	11	172
1990	"Integrated Modeling and Control Based on Reinforcement Learning" by R. S. Sutton	0	44
1991	"Bayesian Model Comparison and Backprop Nets" by David J. C. Mackay	1	38
1992	"Reinforcement Learning Applied to Linear Quadratic Regulation" by Steven J. Bradtke	6	73
1993	"Supervised Learning from Incomplete Data via an EM Approach" by Zoubin Ghahramani, Michael Jordan	12	246
1994	"Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems" by Tommi Jaakkola, Sizarad Singhal, Michael Jordan	10	178
1995	"EM Optimization of Latent-Variable Density Models" by Chris Bishop, M. Svenson, Christopher Williams	1	30
1996	"Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing" by V.Vapnik, Steven Golowich, Alex Smola	2	610 (13364)
1997	"EM Algorithms for PCA and SPCA" by Sam Roweis	1	267

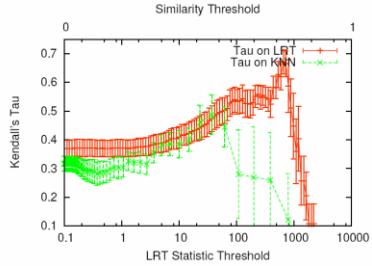
Influence Ranking

$$\tau = \frac{2 \cdot \text{number of concordant pairs}}{\text{total number of pairs} \; - \; \text{number of tied pairs}}$$

τ	LRT	SIM
NIPS	0.4163	0.3686
HEPTH	0.3549	0.3190

Per-Doc Citations	
Predicted	Actual
Doc #1	Doc #1
Doc #2	Doc #3
Doc #4	Doc #4
Doc #3	Doc #2
Doc #5	Doc #5

LRT Statistic Threshold



Document #1	
Score	> 10 ?
500.3	yes
80.3	yes
2.5	no

Document #2	
Score	> 10 ?
294.2	yes
8.4	no
1.1	no

Summary

For collections without a citation graph:

- **Model of influence between documents**
- **Method to construct an influence graph**
- **Method to identify the most influential documents**

Further Questions:

- **Efficiency (all pairs)**
- **Identify novelty**
- **Provide descriptive summaries of ideas**
- **Segmentation of documents**
- **What other things to do with the influence graph?**