MLSS Taipei 2006 Workshop

Identifying Temporal Patterns and Key Players in Document Collections

Rich Caruana, Thorsten Joachims,
Johannes Gehrke, Benyah Shaparenko

Cornell University

{caruana,tj,johannes,benyah}@cs.cornell.edu

Introduction and Goals

- **Identify Development of Topics**
  - What are the key topics in a collection of documents and how did their popularity change over time?
- **Identify Influential Documents**
  - Which documents introduced new ideas that had a large impact?
- **Identify Influential Authors**
  - Which authors have the largest influence on the development of topics?

Key Contribution

- **Find influential documents and authors *without* using citation analysis**
  - Methods based on document text
  - Only metadata is the author name for who wrote which document
  - Wider applicability: news, blogs, email, etc.

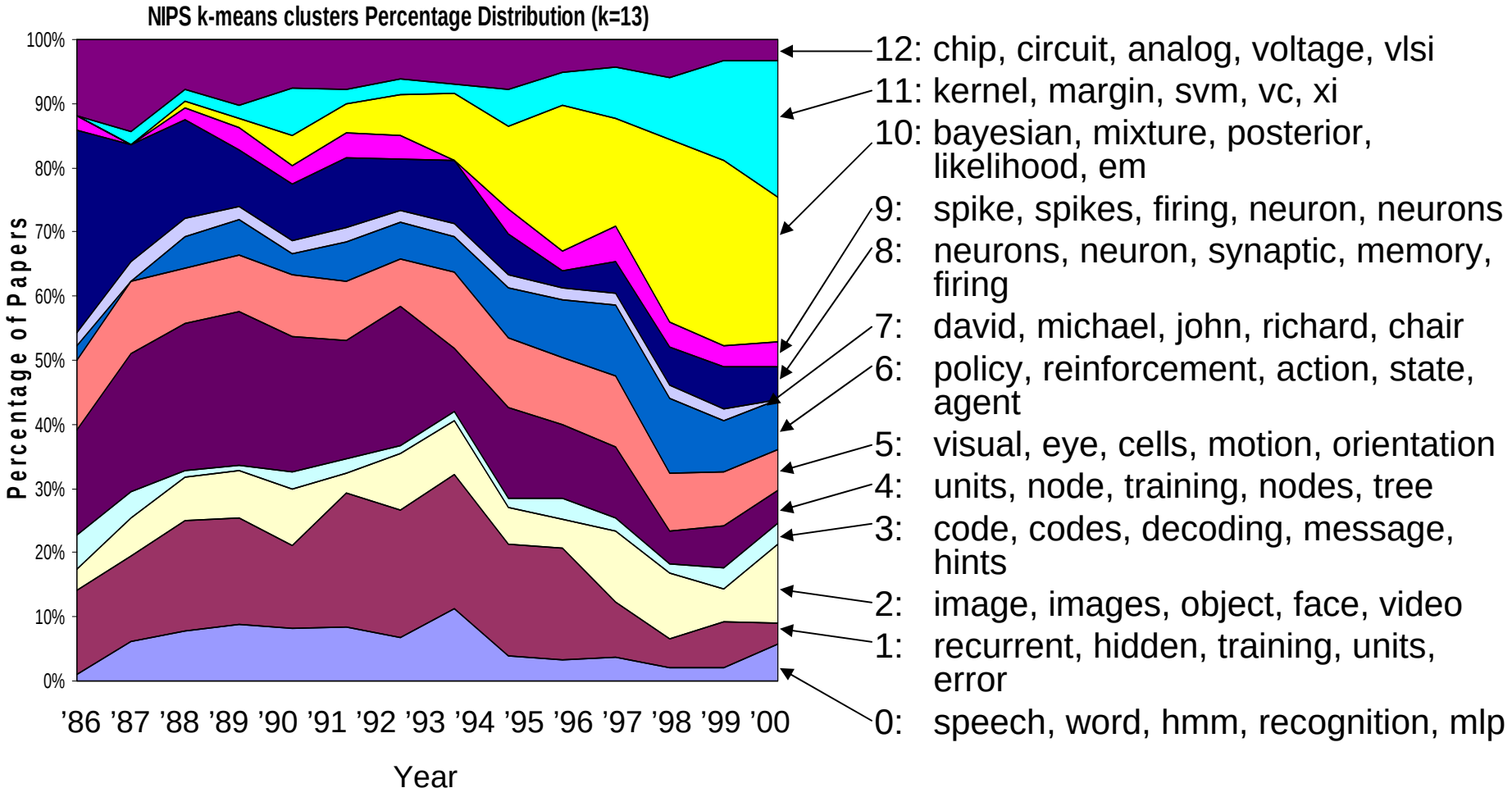
Temporal Cluster Histograms: Goals

- **What are the main topics in a collection?**
  - Identify key topics.
  - What proportion of documents is in each topic?
- **How do topics develop?**
  - What are new emerging topics?
  - Which topics are fading?
  - When did particular topics peak in popularity?

Temporal Cluster Histograms: Method

- **Document Assumptions**
  - Text, Time-stamped, Document dependencies
- **Testbed**
  - 1955 NIPS documents from 1987 – 2000
  - Vector space TFIDF representation
- **K-means Clustering**
  - Cosine distance metric on TFIDF text vectors
  - K = 7, 13, 30
  - 10 runs, select run with least squared error

Temporal Cluster Histograms: Results



Related Work (Topics)

- **Temporal Topic/Trend Detection**
  - Topic Detection and Tracking (TDT) Studies (Allan et al. ’98)
  - ICA (Kolenda et al. ’01)
  - Burst detection (Kleinberg ’02)
  - Timelines (Swan ’00)
  - Efficient, formal models (Guha et al. ’05)
  - Thread life cycle (Mei et al. ’05)

Influential Documents: Goals

- **Identify "leading" papers**
  - Which documents introduce new ideas?
  - Which documents most influence future work?
- **Constraints**
  - Analysis not limited to scientific papers

    → Must work without citation data
  - Only document text and timestamp may be used

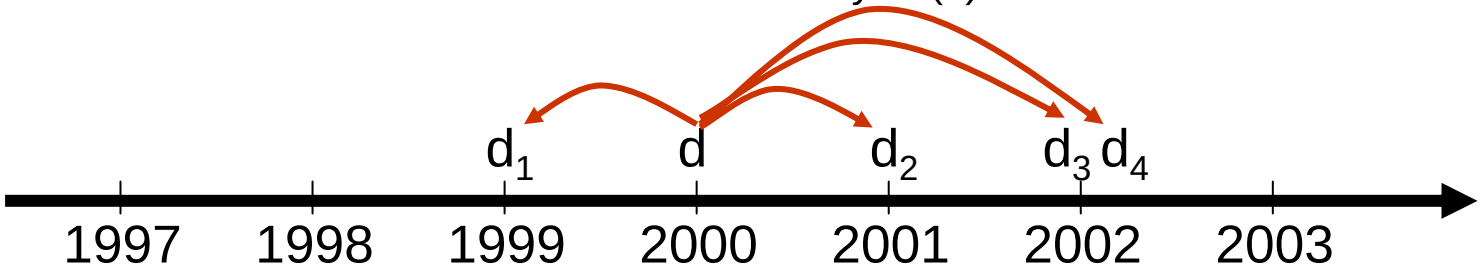
Related Work (Influence)

- **Influential Documents and Authors**
  - Bibliometrics (McGovern et al. ’03, Osareh ’96, White ‘04)
  - Hubs/Authorities (Kleinberg ’99)
  - PageRank (Page et al. ’99)
  - Impact factors (Garfield ’03)

Influential Documents: Method

- **Document Lead/Lag Index**
  - Basic assumption: Document dependencies are based on the terminology of documents.
  - Find the k nearest neighbors (cosine distance)
  - Raw lead/lag score: $LL_{raw}(d)$ = # later – # earlier
  - Scaled lead/lag score (avoid edge effects):

$$LL_{norm}(d) = LL_{raw}(d) - AVG_{d' \in year(d)}(LL_{raw}(d'))$$



Influential Documents: Results

Score	Year	Cites	Paper Title and Authors
1.167	1996	128	“improving the accuracy and speed of support vector machines” by chris j.c. burges, b. schoelkopf
1.128	1999	17 (466)	“using analytic qp and sparseness to speed training of support vector machines” by john c. platt
0.986	1999	18	“regularizing adaboost” by gunnar raetsch, takashi onoda, klaus-robert mueller
0.953	1996	41 (3711)	“support vector method for function approximation, regression, and signal processing” by v. vapnik, s. golowich, a. smola
0.945	1998	27	“training methods for adaptive boosting of neural networks” by holger schwenk, yoshua bengio
0.945	1997	3	“modeling complex cells in an awake macaque during natural image viewing” by william e. vinje, jack l. gallant
0.934	1998	17	“em optimization of latent-variable density models” by chris bishop, markus svensen, chris william
0.934	1995	584	“a new learning algorithm for blind signal separation” by s. amari, a. cichocki, h. h. yang

Influential Authors: Goals

- **Identify "leading" authors**
  - Who are the major players on the scene?
  - Which authors most influence future work?
- **Constraints**
  - Analysis not limited to scientific papers

    → Must work without citation data
  - Only document text and timestamp may be used

Influential Authors: Method

- **Author Lead/Lag Index**
  - Assume author *a* has documents $d_1,\ldots,d_n$
  - Compute scaled lead/lag score for each document and average these scores

    $LL_{norm}(a) = 1/n\ (LL_{norm}(d_1)+\ldots+ LL_{norm}(d_n))$
  - Compute variance *v* of $LL_{norm}(a)$ and rank by

    $LL_{norm}(a) - 2 * sqrt(v / n)$
  - Use smoothing to avoid small sample artifacts

Influential Authors: Results

Author	Rank	Papers	Citations
jordan, michael i.	0.037	27	3400
smola, alex	-0.004	13	1499
scholkopf, b.	-0.022	10	1697
atkeson, christopher g.	-0.06	10	693
williams, christophe k.i.	-0.067	16	2363
sejnowski, terrence j.	-0.069	46	2210
hinton, geoffrey e.	-0.075	27	3774
jaakkola, tommi	-0.091	10	930
miller, kenneth d.	-0.106	11	1634
coon, d. d.	-0.112	21	14
bengio, yoshua	-0.131	18	929
saad, david	-0.133	11	573
bialek, william	-0.135	11	520

Conclusions

- Propose problem: Analyze temporal trends without using citation data
- Temporal cluster histograms concisely depict how popularity changes over time.
- Document (author) lead/lag index identifies important, influential documents (authors) without using citation analysis.

Future Directions

- Better methods for identifying influential documents and authors
  - Simple methods work
  - More principled methods probably do better
- Associate influential documents with cluster formation
- Clustering algorithms that directly capture the splitting and merging of clusters