# Publication Venue Recommendation based on Paper Abstract

A.Bartoli, E.Medvet, G.Piccinin
*University of Trieste, Italy*

MACHINE LEARNING LAB

# The Problem

- Given a scientific **paper** $p$
- Recommend an ordered list of “suitable” **publication venues** $v_1, v_2, \ldots, v_N$

# Motivation

- Everyone knows “the” top-level conferences
- Choosing the right venue for early / exploratory work is not easy
- More than 2000 CS conferences
  - Broad spectrum between generalist vs specific
  - How to have high precision and recall ?
- Some emerging proposals
- ...and to be frank: a nice and challenging problem...

# Novelty and Advantages

- Recommendation based solely on **title** and **abstract**
- Unlike earlier proposals based on **full-text** and **reference list**

- May be used **earlier** in the authoring process
- KB much **simpler** to build and maintain

# Our work in a nutshell

- Three approaches
- Assessed on 58000 papers / 300 conferences from Microsoft Academic Search

- Results aligned with existing state of the art
- ...but with much less info from paper and KB !

# Approach 1: Cavnar-Trenkle

Assumption:

- Each venue has a specific **language profile**
  - Profile: n-gram list sorted by frequency ($n \leq 5$)

Recommend:

- Venues with language profiles “closest” to the examined paper

# LDA in a nutshell (I)

INPUT

- Collection of papers
- Predefined #topics

OUTPUT

- Each **topic** is a **mix of words**:
  word vector (prob. of finding that word in “this” topic)
- ...

# Example topics

| Topic | 4 most probable words | | | |
|---|---|---|---|---|
| 1 | data | analysi | mobil | network |
| 2 | system | network | mobil | comput |
| 3 | system | process | analysi | comput |
| 4 | network | sensor | wireless | system |
| 5 | network | data | algorithm | perform |
| 6 | comput | network | servic | perform |

- A topic “**is**” just a word probability vector
- Topics are discovered (generated) **automatically**
- We arbitrarily set #topics=20

# LDA in a nutshell (II)

INPUT

- Collection of papers
- Predefined #topics

OUTPUT

- Each **topic** is a **mix of words**:
  word vector (element *k* = prob. of being found in topic *k*)
- Each **paper** is a **mix of topics**
  topic vector (element *k* = prob. of being related to topic *k*)

# Approach 2: 2-Step LDA

Assumption:

- Each venue has a **prevalent topic**
  - The one most probable for most papers
- **Sub**-topics are the topics generated with LDA **restricted** to venues with **the same prevalent topic**

Recommend:

- Venues with prevalent topic and sub-topics “closest” to the examined paper

# Approach 3: LDA+Clustering

Assumption:

- Papers may be **clustered** based on their topic mix
- **Sub**-topics are the topics generated with LDA **restricted** to papers from **the same** ~~prevalent topic~~ **cluster**

Recommend:

- Venues with subtopic mix “closest” to the examined paper

# Experimental Evaluation

- 58000 papers / 300 conferences
  from Microsoft Academic Search

- Half training (Knowledge Base)
- Half testing (Recommend)
  - “Half” = same #papers in each conference

- Two repetitions (with different splitting)

# Metric

- Venue-Accuracy@N
  - Prediction is correct when:
    real venue is in the N recommended venues
- Probably excessively (and unnecessarily) severe
  - A paper may fit many conferences

# Baseline

- Random recommender

- Earlier proposals: not really meaningful
  - Very different datasets
    (size, content)
  - Keep in mind: they need **full text** and **reference list**

# Venue-Accuracy@N

| Method | venue-acc.@$N$ (%) $N=3$ | $N=5$ | $N=10$ | Dataset $\lvert A \rvert$ | $\lvert V \rvert$ |
|---|---|---|---|---|---|
| Cavnar-Trenkle | 26.8 | 34.0 | 45.6 | | |
| | | | | 58466 | 300 |
| Random recommender | 1.0 | 1.7 | 3.3 | | |
| [2] ACM | | | | | |
| [2] CiteSeer | | | | | |
| [3] | | | | | |
| [4] | | | | | |

- Simple n-gram profile is indeed effective

# Venue-Accuracy@N

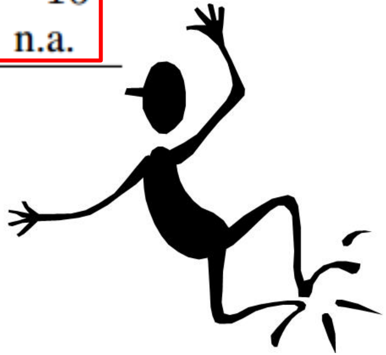
| Method | venue-acc.@$N$ (%) $N=3$ | $N=5$ | $N=10$ | Dataset $\|A\|$ | $\|V\|$ |
|---|---|---|---|---|---|
| Cavnar-Trenkle | 26.8 | 34.0 | 45.6 | 58466 | 300 |
| Two-step-LDA | 3.4 | 3.8 | 4.0 | | |
| LDA+clustering | 16.1 | 21.7 | 33.2 | | |
| Random recommender | 1.0 | 1.7 | 3.3 | | |
| [2] ACM | | | | | |
| [2] CiteSeer | | | | | |
| [3] | | | | | |
| [4] | | | | | |

- Topic models may *(or may not)* be effective
- Many matching criteria

# Earlier proposals

| Method | venue-acc.@$N$ (%) $N=3$ | $N=5$ | $N=10$ | Dataset $\|A\|$ | $\|V\|$ |
|---|---|---|---|---|---|
| Cavnar-Trenkle | 26.8 | 34.0 | 45.6 | 58466 | 300 |
| Two-step-LDA | 3.4 | 3.8 | 4.0 | | |
| LDA+clustering | 16.1 | 21.7 | 33.2 | | |
| Random recommender | 1.0 | 1.7 | 3.3 | | |
| [2] ACM | - | 55.7 | 69.8 | 172 890 | 2197 |
| [2] CiteSeer | - | 23.9 | 29.0 | 35 020 | 739 |
| [3] | 91.6 | - | - | 960 | 16 |
| [4] | - | - | 63.2 | 295 317 | n.a. |

- Relying on **only title and abstract** seem to be enough !



# A Weaker Metric

- **Subdomain-**Accuracy@N
    - Each venue has one or more of 24 subdomains (assigned by Microsoft Academic Search)

    - Prediction is correct when:

      subdomains of real venue and of N recommended venues have non-empty intersection

# Subdomain-Accuracy@N

| Method | sub-domain-acc.@$N$ (%) $N=3$ | $N=5$ | $N=10$ |
|---|---|---|---|
| Cavnar-Trenkle | 54.1 | 61.1 | 70.9 |
| Two-step-LDA | 9.9 | 10.1 | 10.2 |
| LDA+clustering | 47.3 | 56.5 | 68.9 |
| Random recommender | 14.3 | 22.6 | 40.1 |

- “Similar” conclusions

# Thanks for your attention

MACHINE LEARNING LAB

University of Trieste, Italy

http://machinelearning.inginf.units.it

HIGH-FREQUENCY SHAPE AND ALBEDO FROM SHADING USING NATURAL IMAGE STATISTICS

We relax the long-held and problematic assumption in shape-from-shading (SFS) that albedo must be uniform or known, and address the problem of "shape and albedo from shading" (SAFS). Using models normally reserved for natural image statistics, [. . . ]

AN EFFICIENT COMMUNITY DETECTION METHOD USING PARALLEL CLIQUEFINDING ANTS

Attractiveness of social network analysis as a research topic in many different disciplines is growing in parallel to the continuous growth of the Internet which allows people to share and collaborate more Nowadays detection of community structures [. . . ]

FASTER EXPLICIT FORMULAS FOR COMPUTING PAIRINGS OVER ORDINARY CURVES

We describe efficient formulas for computing pairings on ordinary elliptic curves over prime fields. First, we generalize lazy reduction techniques, previously considered only for arithmetic in quadratic extensions, to the whole pairing computation, including toweri arithmetic. [. . . ]

1. *Computer Vision and Pattern Recognition*
2. Storage and Retrieval for Image and Video Databases
3. International Conference on Computer Vision

1. International Conference on Weblogs and Social Media
2. Recent Advances in Intrusion Detection
3. IEEE INFOCOM

(*IEEE Congress on Evolutionary Computation*)

1. Pairing-Based Cryptography
2. International Parallel and Distributed Processing Symposium/International Parallel Processing Symposium
3. International Conference on Computational Science

(*Theory and Application of Cryptographic Techniques*)