Instituto Superior Técnico
Universidade Técnica de Lisboa

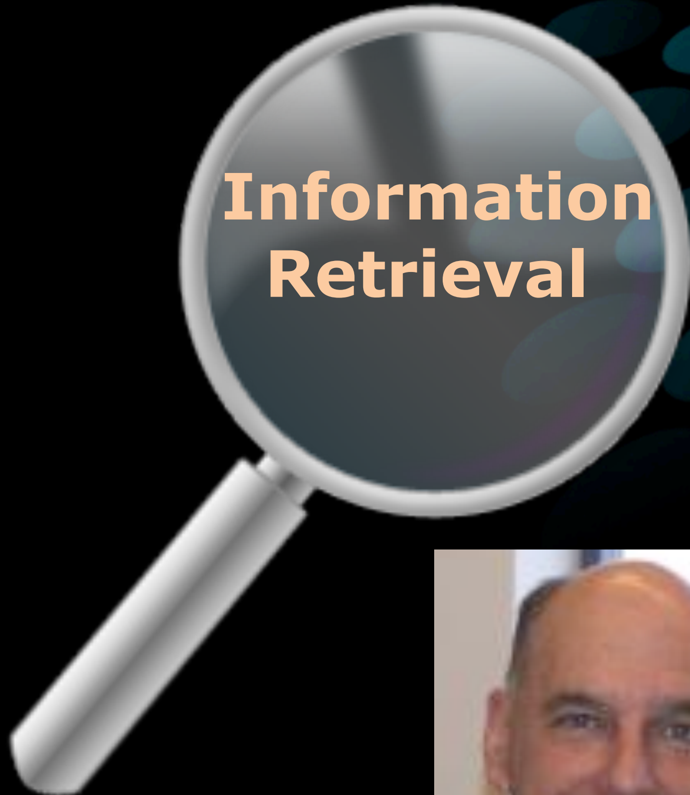
Learning to Rank Academic Experts

Catarina Moreira, Pável Calado and Bruno Martins

Outline

- ✓ Introduction
- ✓ Related Work
- ✓ Learning to Rank
- ✓ Features
- ✓ Algorithms
- ✓ Dataset
- ✓ Experimental Results

Expert Finding



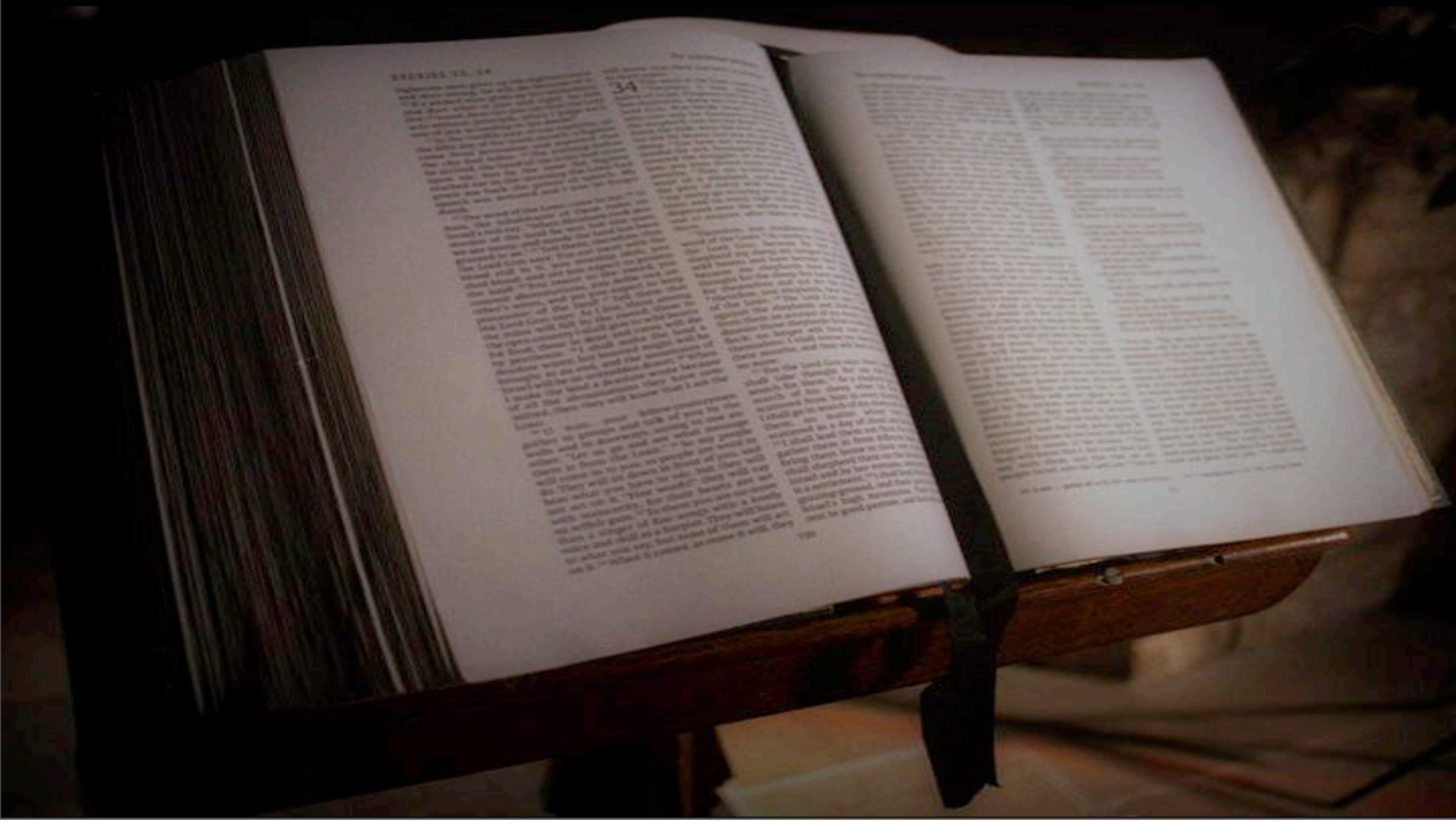
Why Expert Finding?

Too many documents

Information is dispersed

Need answers quickly

Related Work



Candidate Centric Approach

1. Gather documents associated to a candidate
2. Merge documents into a single profile document
3. Rank the profile according to the query

Document Centric Approach

1. Gather documents containing query topics

2. Uncover candidates and rank them

Problems?

Generative Probabilistic Models

Simple heuristics

Heuristics do not reflect expertise

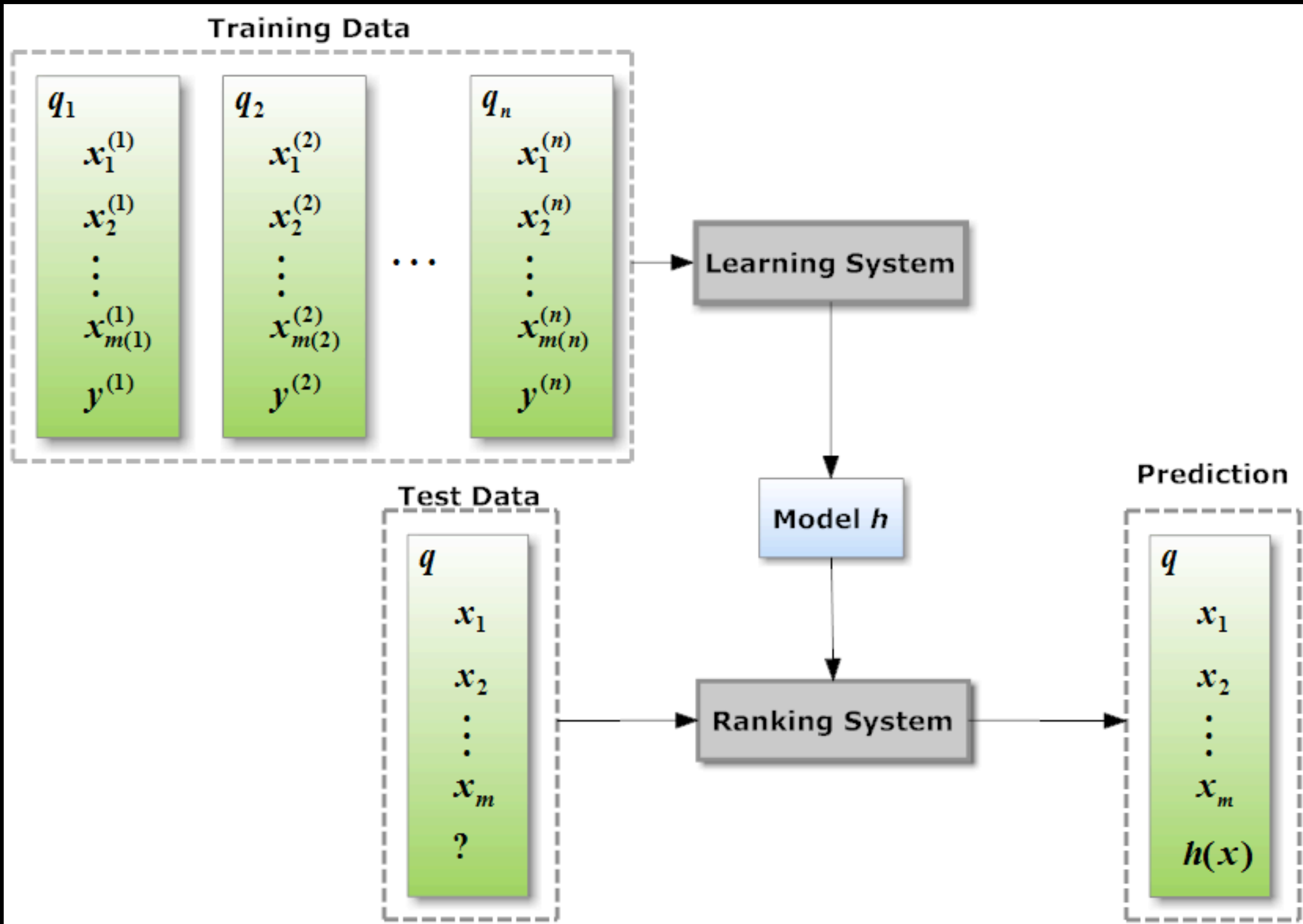
Only based on textual contents

Our Approach

A set of features to estimate expertise

Features combined in a **learning to rank framework**

Learning to Rank (L2R)



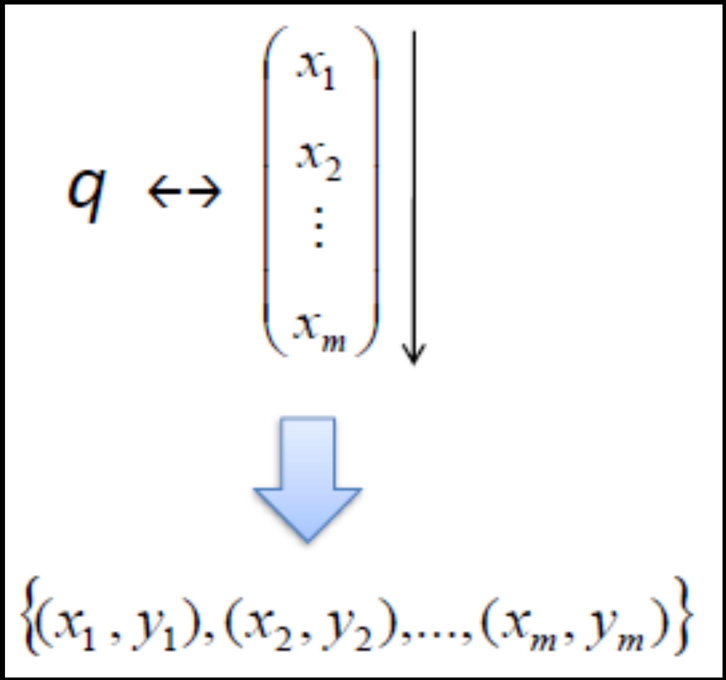
L2R Approaches

- Pointwise
- Pairwise
- Listwise

L2R Pointwise Approaches

Use feature vectors for each individual **<q, x>**

Goal: directly support the application of existing algorithms of regression or classification



L2R Pairwise Approaches

Use feature vectors for each

pair **<q, x1, x2>**

Goal: minimize number of

misclassified document pairs

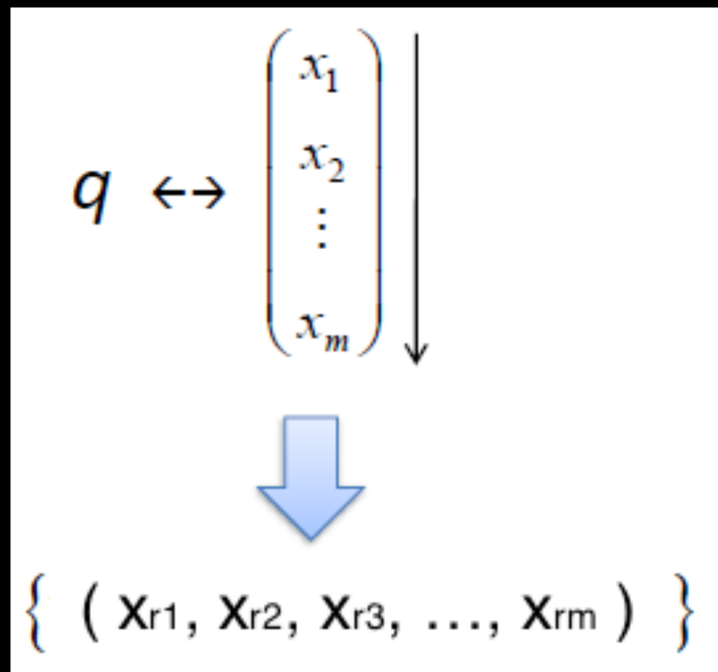
$$q \leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$\left\{ \begin{matrix} (x_1, x_2, +1), (x_2, x_1, -1), \ldots, \\ (x_2, x_m, +1), (x_m, x_2, -1) \end{matrix} \right\}$$

L2R Listwise Approaches

Use feature vectors for the list **<q, x1, x2, ..., xm>**

Goal: optimize an Information Retrieval evaluation metric



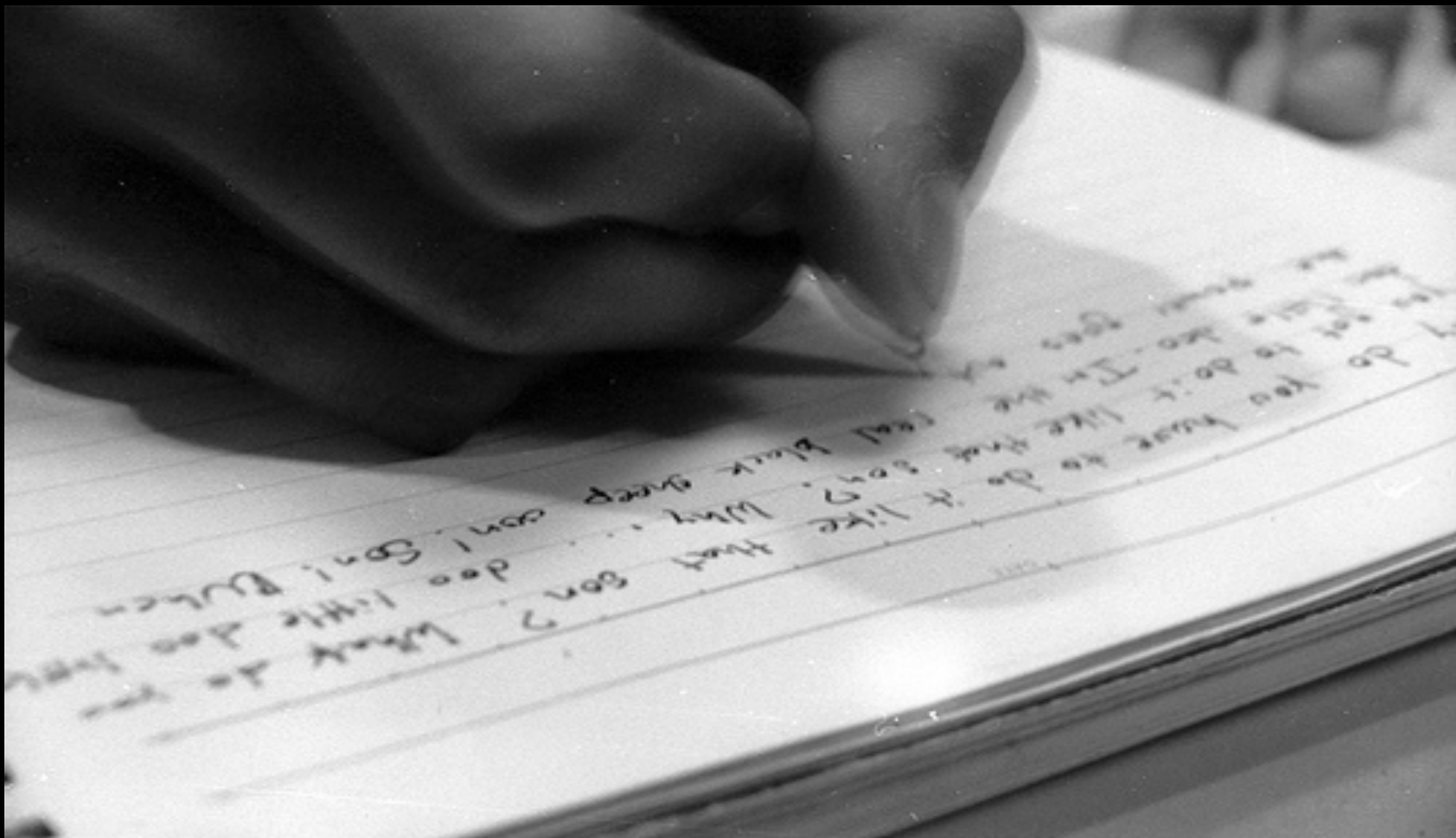
Features

Textual Similarities

Profile Information

Graph Structure

Textual Features



Textual Features

TF

$$TF_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \frac{Freq(i, d_j)}{|d_j|}$$

IDF

$$IDF_q = \sum_{i \in Terms(q)} \log \frac{|D|}{f_{i,D}}$$

BM25

$$BM25_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \log \left(\frac{N - Freq(i) + 0.5}{Freq(i) + 0.5} \right) \times \frac{(k_1 + 1) \times \frac{Freq(i,d_j)}{|d_j|}}{\frac{Freq(i,d_j)}{|d_j|} + k_1 \times (1 - b + b \times \frac{|d_j|}{\mathcal{A}})}$$

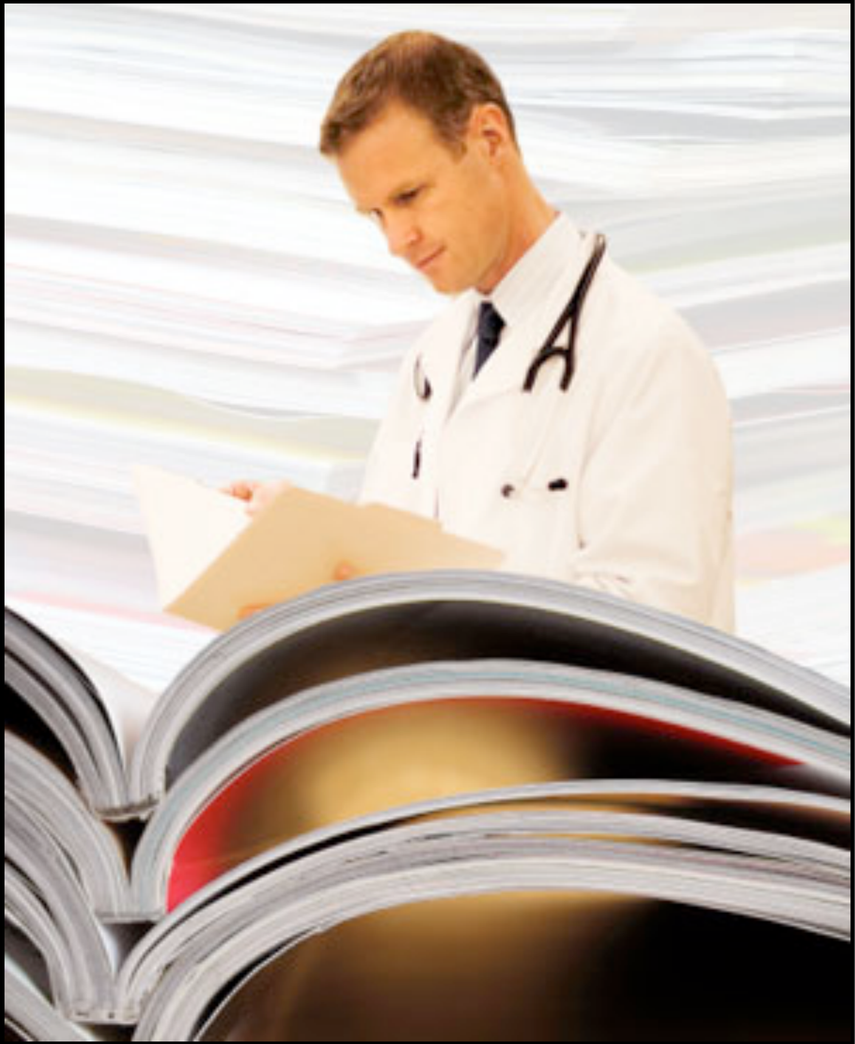
Profile Features

Profile Features

Number of Publications

Years Between Publications

Number of Articles



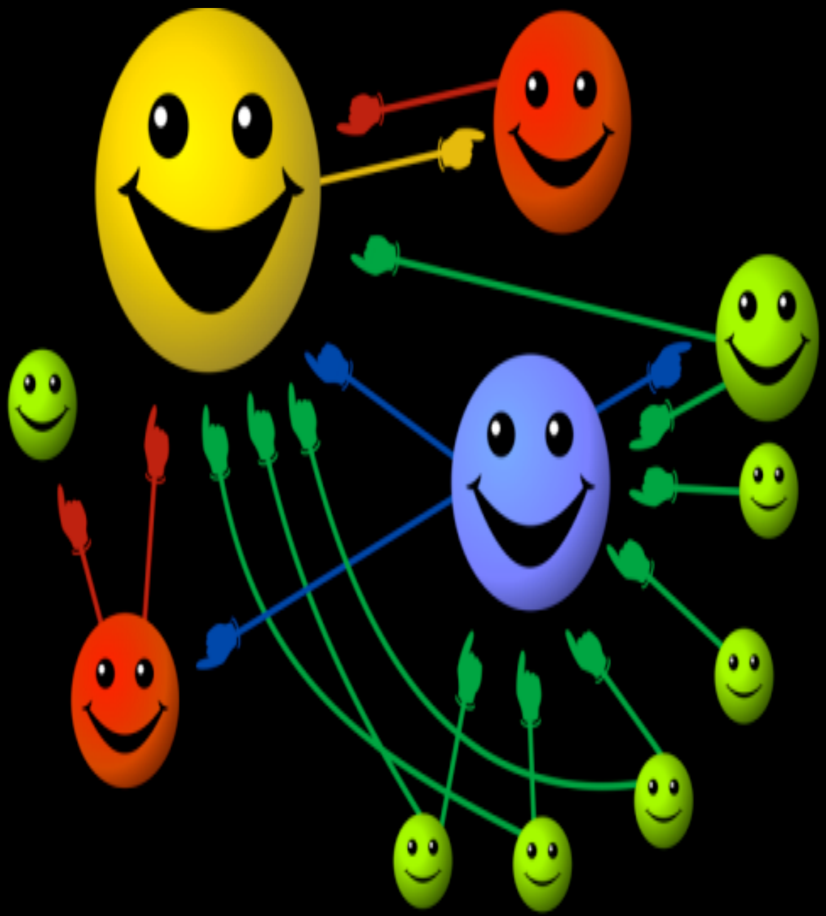
Graph Features

Graph Features

Citations Graphs

Co-authorship Graphs

Academic Indexes



Academic Indexes Measure Scientific Impact!

Academic Indexes

H-Index

G-Index

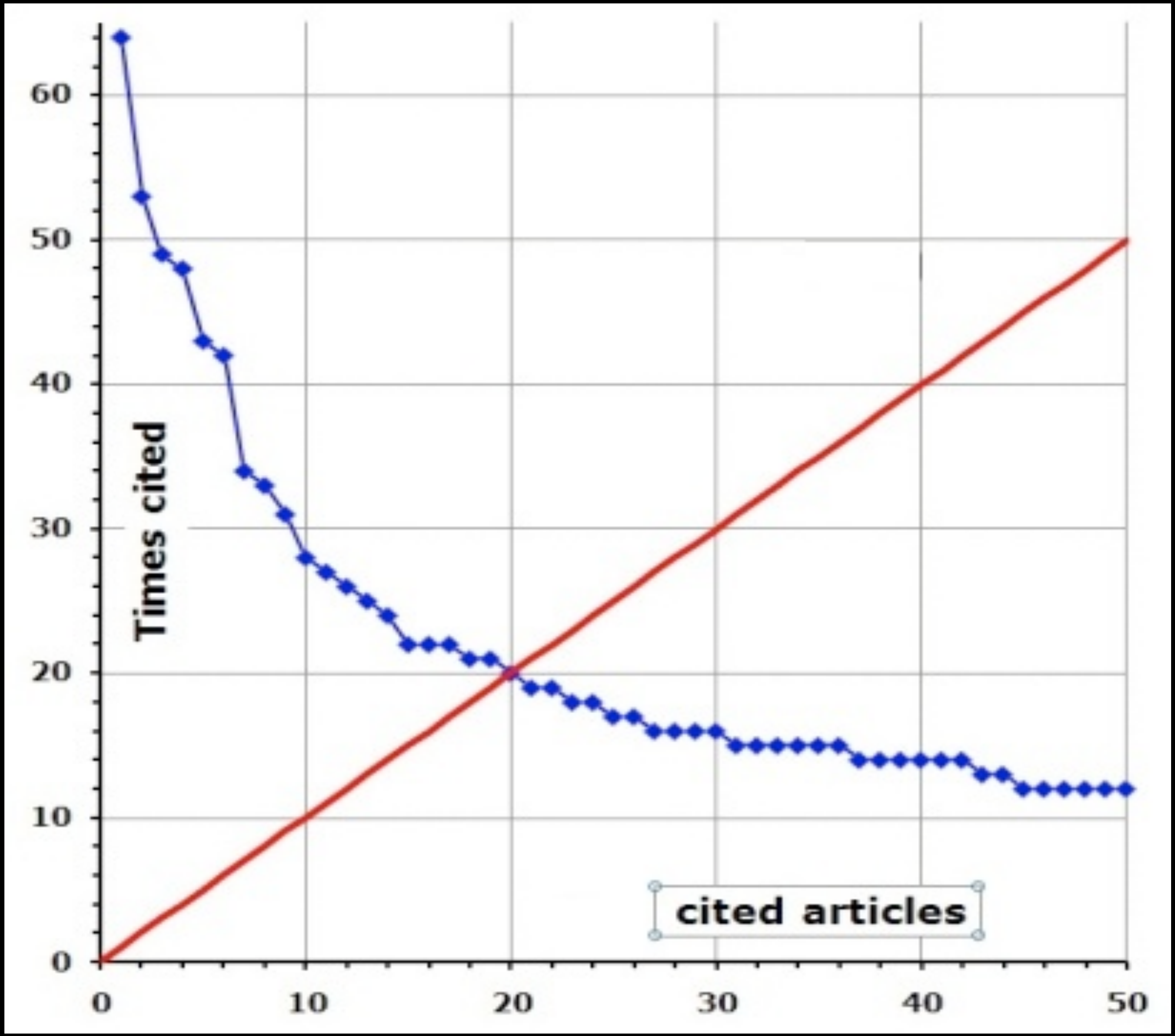
A-Index

H Index

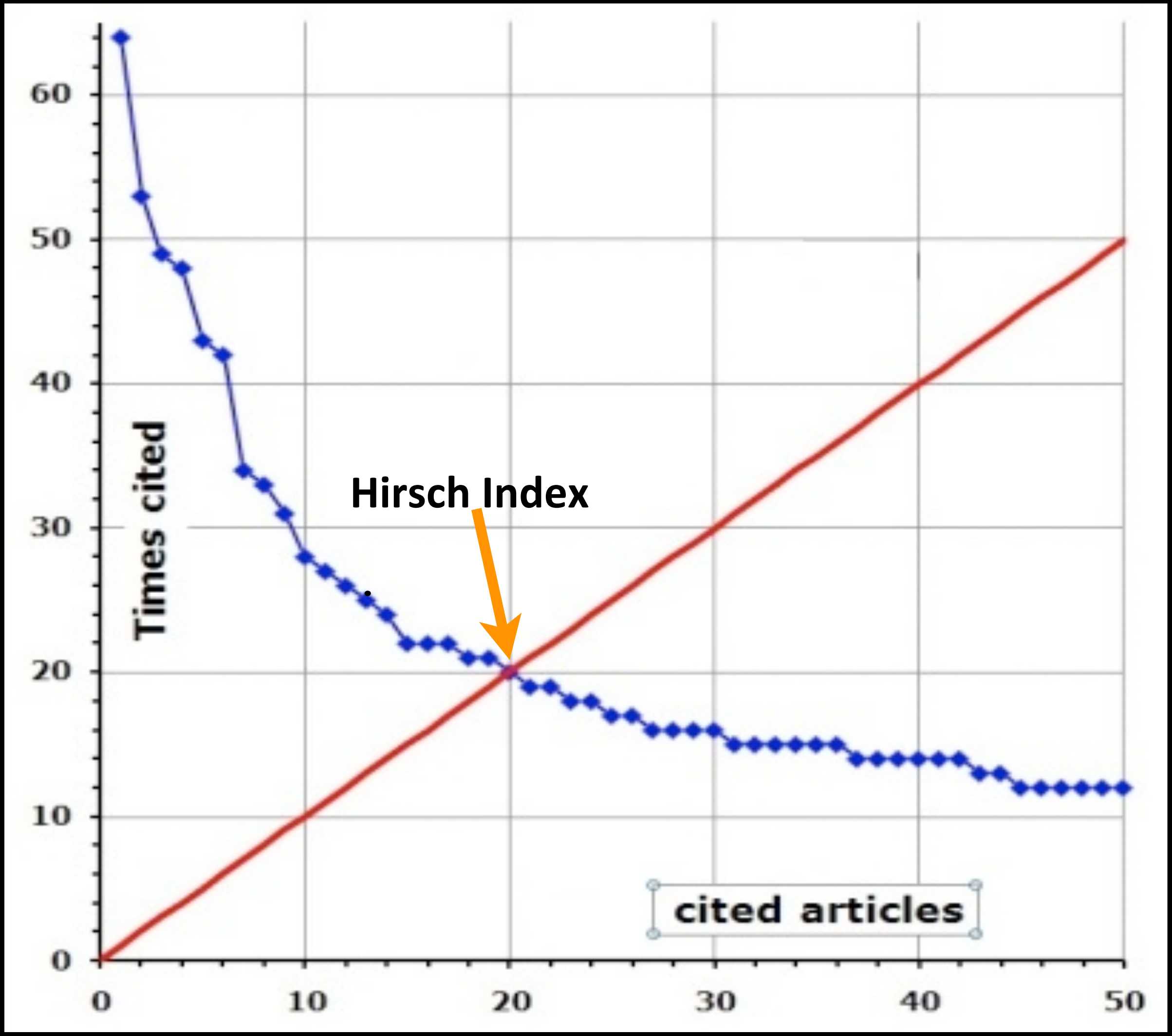
A given author has a Hirsch Index of *h*, if *h*

of his *N* papers have at least *h* citations each

H Index - Example



H Index - Example



G Index

Is the largest number such that the top *g* papers received on average at least *g* citations each

a Index

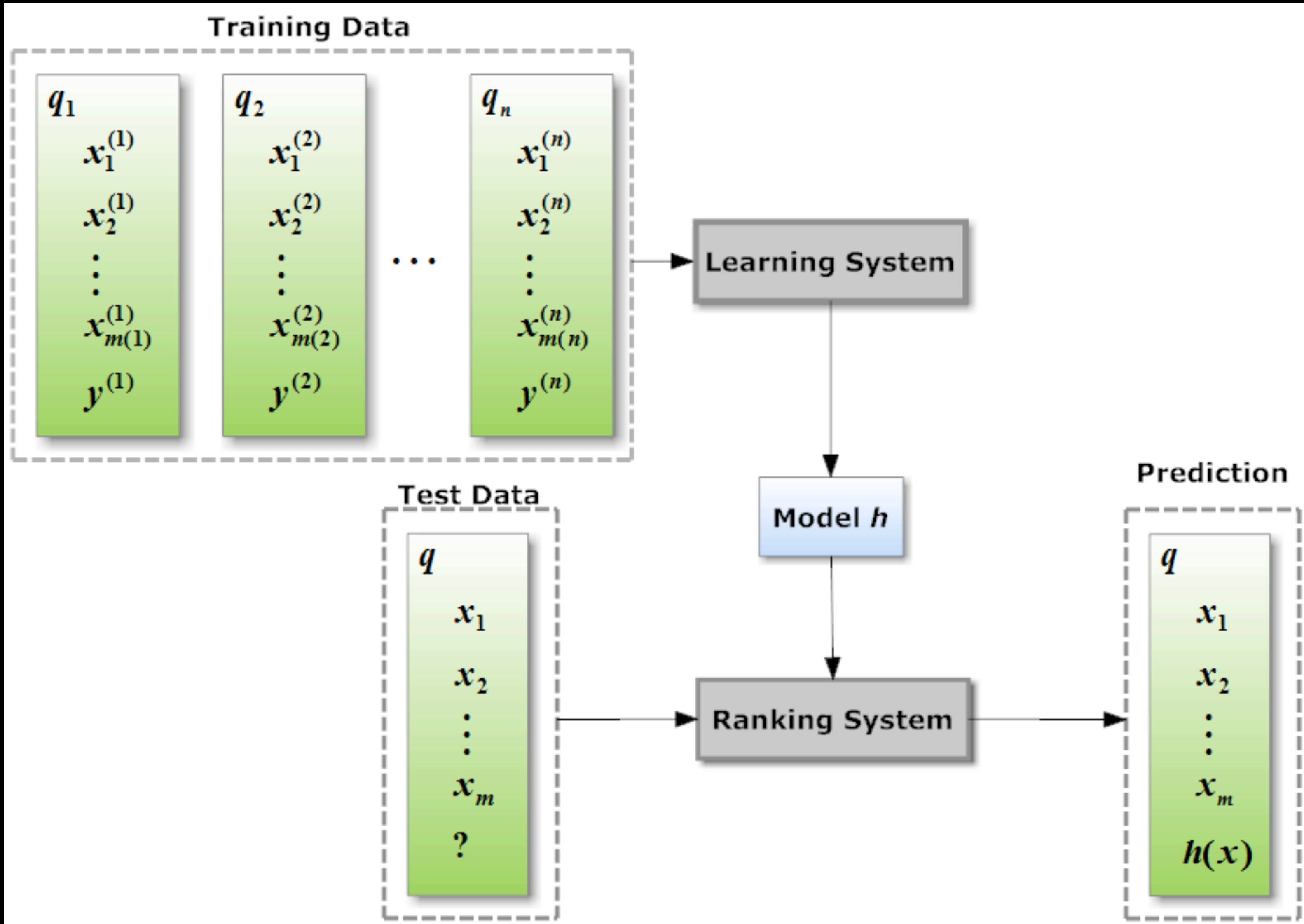
Measures the magnitude of the most influential papers of a given author

$$a = N_{c,tot}/h^2$$

First work using academic indexes

to estimate Expertise!

Learning to Rank (L2R)



L2R Algorithms Tested

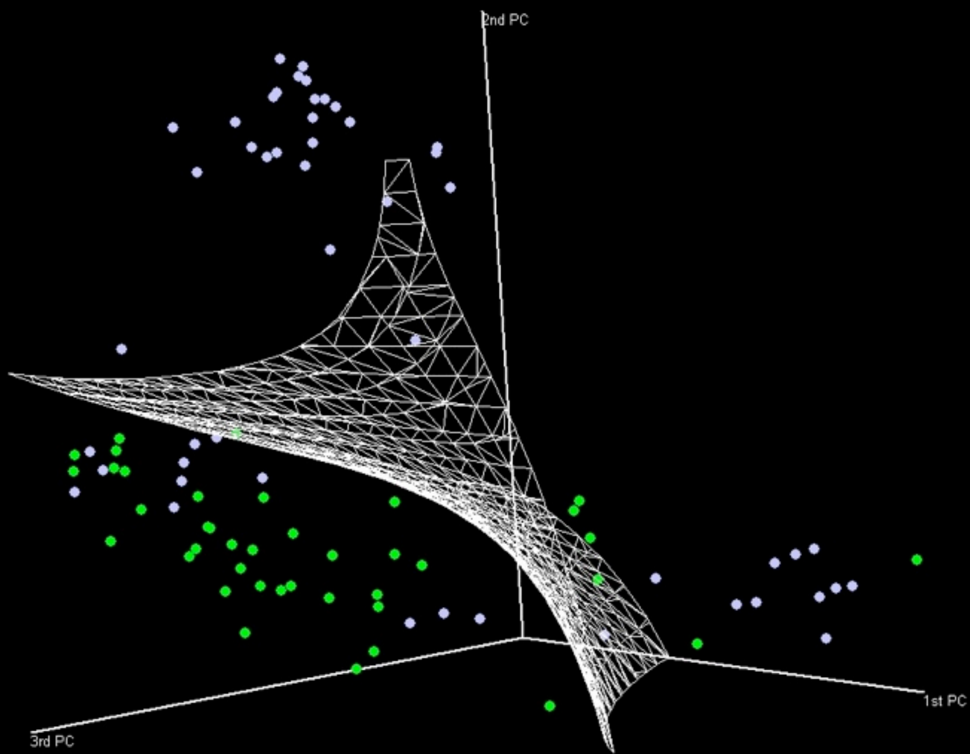
Based on the formalisms of Support Vector Machines:

- **SVMmap** (Y. Yue and T. Finley)
- **SVMrank** (T. Joachims)

Support Vector Machines

Basic idea:

Construct an N-dimensional hyperplane to separate data points.



SVMmap

Optimizes MAP by minimizing a loss function which measures the difference between a perfect ranking and the performance of an incorrect ranking

$$\min \frac{1}{2}||w||^2 + \frac{C}{m}\sum_{i=1}^{m}\xi^{(i)}$$

$$\text{s.t. } \forall y^{c(i)} \neq y^{(i)}, w^T\Psi(y^{(i)}, x^{(i)}) >= w^T\Psi(y^{c(i)}, x^{(i)}) + 1 - AP(y^{c(i)}) - \xi^{(i)}$$

SVMrank

Constrains the default SVM optimization problem to perform to minimization of misclassified **pairs** of experts

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \sum_{u,v:y_{u,v}^{(i)}} \xi_{u,v}^{(i)}$$

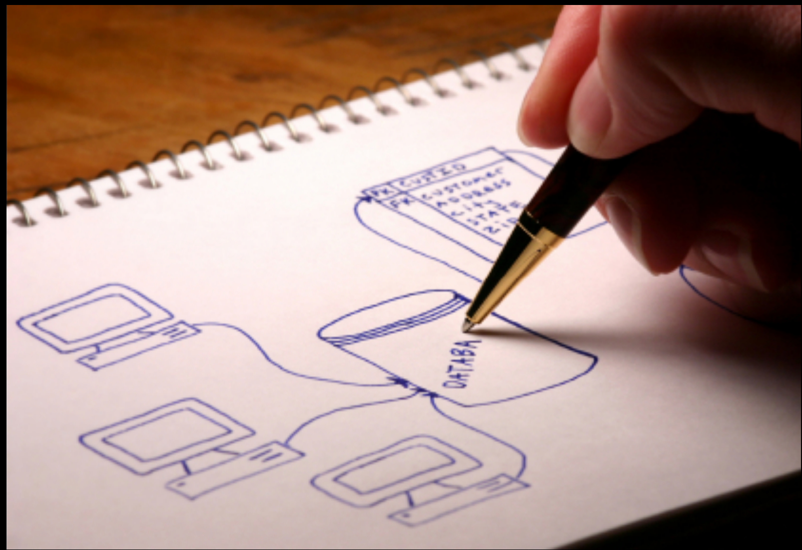
$$\text{s.t. } w^T(x_u^{(i)} - x_v^{(i)}) >= 1 - \xi_{u,v}^{(i)} \text{ ,}$$

$$\text{if } \; y_{u,v}^{(i)} = 1, \; \xi_{u,v}^{(i)} >= 0 \text{ , } \; i = 1, \ldots, n$$

Dataset

DBLP Computer Science Bibliography

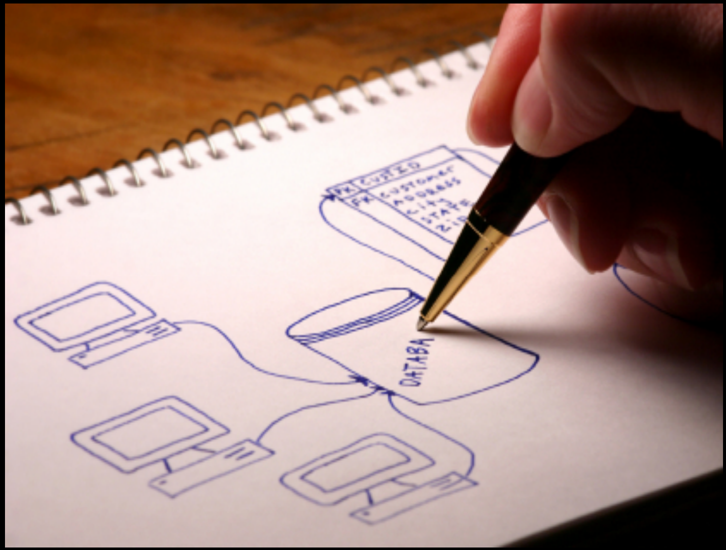
- Covers journal and conference publications
- Contains publication's abstracts
- Contains citation links



Dataset for Validation

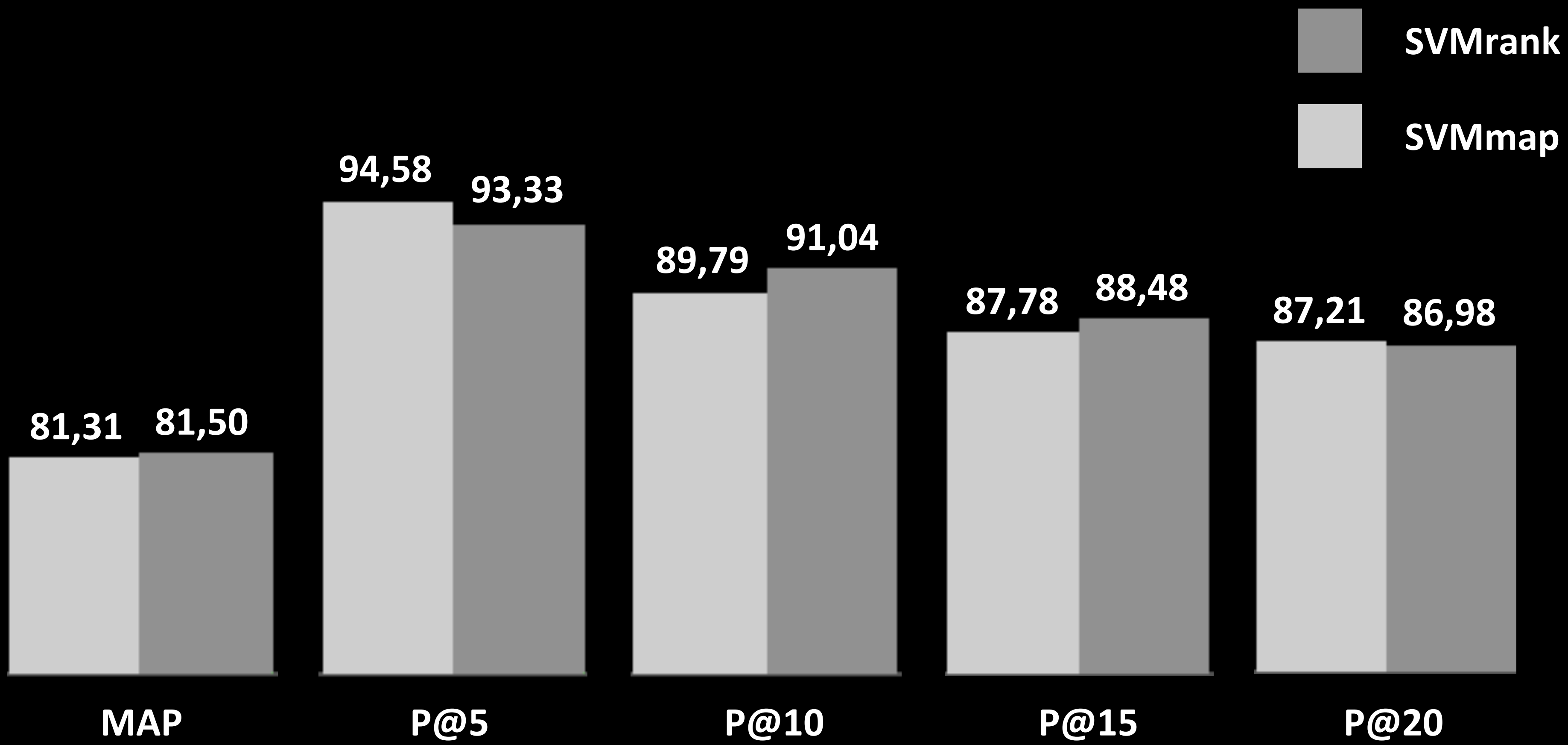
Arnetminer

- Contains a set of people considered experts
- Contains 13 different query topics
- Based on people from Program Committees of important conferences



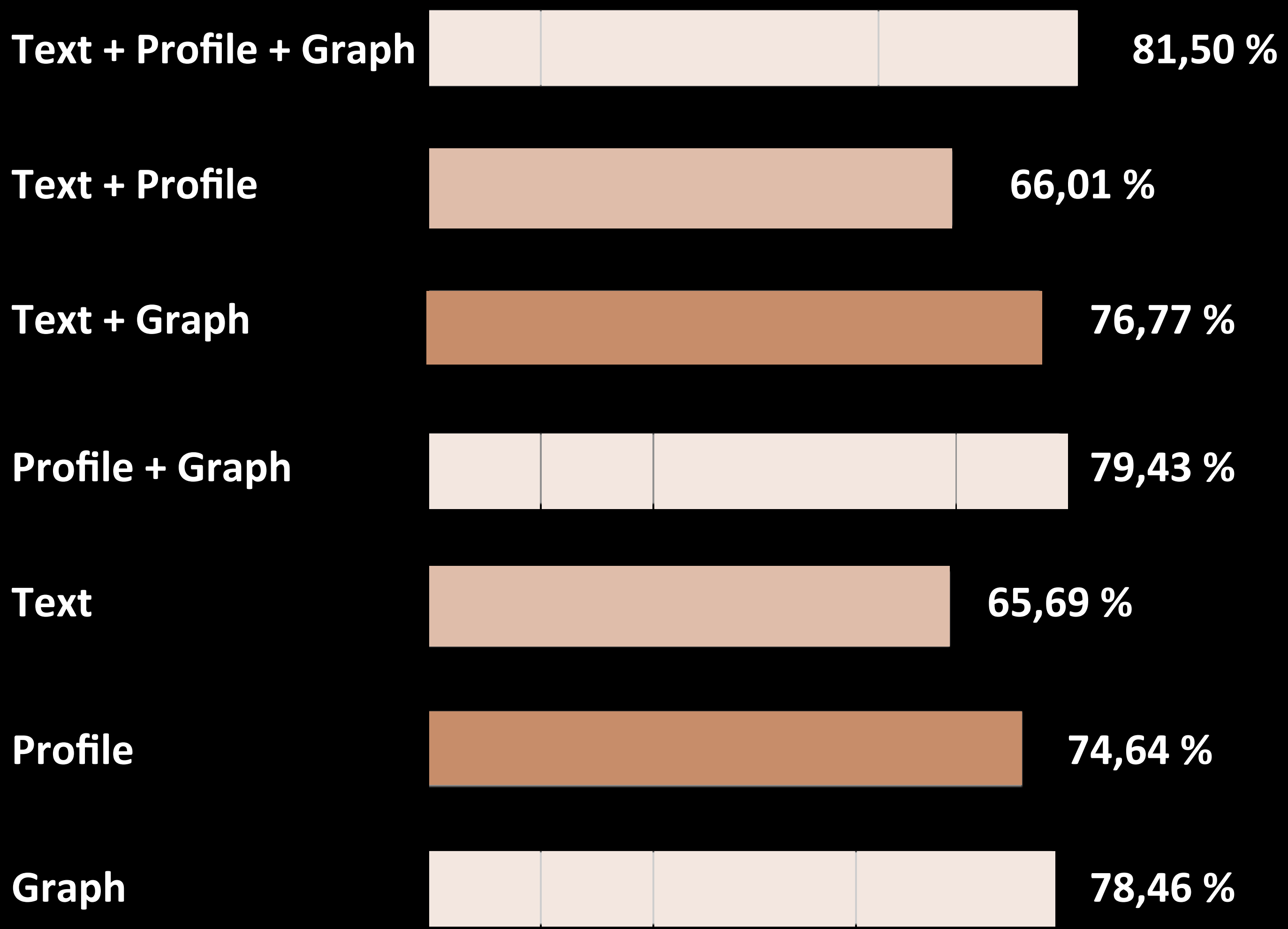
Experimental Results

SVMmap vs SVMrank (%)



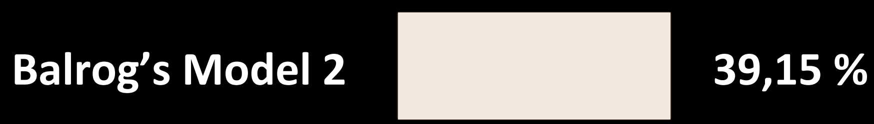
Impact of the Features?

Impact of the Features?



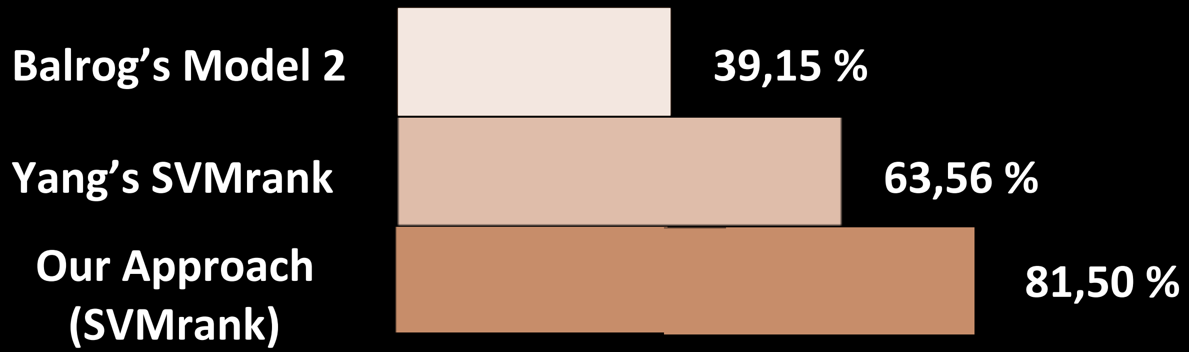
SVMrank Impact in the State of the Art?

SVMrank vs State of the Art (MAP)



SVMrank vs State of the Art (MAP)

SVMrank vs State of the Art (MAP)



Thank You!

Questions?