Ten Years of Wisdom & Beyond

Ricardo Baeza-Yates

Agenda

• WSDM’s Birth
• Analysis of Ten Years
• Current & Future Challenges
  • Contextual Multilingual Semantic Search
  • Bias in the Web
WSDM’s Birth and Governance (1/2)

- **May 12, 2006**: e-mail from Ziv Bar-Yossef (Google Haifa) and Junghoo Cho (UCLA)

- **Initial SC**: Rakesh Agrawal (Microsoft)  
  Ricardo Baeza-Yates (Yahoo)  
  Krishna Bharat (Google)  
  Andrei Broder (Yahoo)  
  Soumen Chakrabarti (IIT Bombay)  
  Jon Kleinberg (Cornell)  
  Rajeev Motwani (Stanford)  
  Prabhakar Raghavan (Yahoo)

- **May 25, 2006**: meeting at WWW 2006 in Edinburgh.  
  Most of the SC plus Sue Dumais (Microsoft Research), Ravi Kumar & Andrew Tomkins (Yahoo), Ronny Lempel (IBM), Yoelle Maarek (Google), Mark Manasse (Microsoft), Marc Najork (Microsoft), and Torsten Suel (Polytechnic University)

- **SIGs proposed**: SIGACT, SIGIR, SIGKDD, SIGMOD, and SIGWEB

- **Plan A**: September 2007, with the advice from ACM President, Stu Feldman, to request the SIGWEB sponsorship.

- **Early 2007 conversations**: SIGWEB (Ronny) and SIGMOD (Andrew), plus SIGIR & SIGKDD (myself).

- Asked to be conference chair for the first conference with Andrei and Soumen as PC-Chairs, Ravi as treasurer, and Utkarsh Srivastava of Stanford as local organizer

WSDM’s Birth and Governance (2/2)

- **Our plans changed during my negotiation with SIGIR (Jamie Callan), as the conference was between SIGIR and CIKM, two SIGIR sponsored conferences**

- **Plan B**: second week of February 2008, with the second conference planned in Barcelona.  
  As I was the natural chair for 2009, I stepped down for 2008, being replaced by Marc Najork

- **Driving team**: Andrei, Andrew, Junghoo, Marc, Ravi, Ricardo, Ronny, Soumen & Zvi

- **April 2007**: formal support of SIGKDD and SIGMOD  
  However SIGIR wanted to have a steering committee with its own representative and clear governance rules

- **May 2007**: Andrei, Andrew, Marc, Ravi, Ricardo, Soumen, and Ziv  
  discussed the governance rules at WWW 2007 in Banff, Canada

- **Steering committee rules**: 8 members, 4 of them representing the sponsoring SIGs, balancing industry and academia members, with periods of 4 years

- With the new rules, SIGIR and SIGWEB became sponsors too
WSDM’s Formal Steering Committee

June 2007:
- Rakesh Agrawal (Microsoft, KDD rep.)
- Ziv Bar-Yossef (Google)
- Monika Henzinger (EPFL, Google)
- Rajeev Motwani (Stanford)
- Ricardo Baeza-Yates (Yahoo, SIGIR rep., chair)
- Soumen Chakrabarti (IIT Bombay)
- Jon Kleinberg (Cornell)
- Prabhakar Raghavan (Yahoo)

January 2010: Marc Najork (Microsoft) replaced Prabhakar, and soon after
Hector Garcia-Molina (Stanford) joined representing SIGMOD

May 2013: SC Renewal, almost 2 years late!
- Ricardo Baeza-Yates (Yahoo, SIGIR rep., chair)
- Andrei Broder (Google)
- Nick Koudas (Univ. of Toronto, SIGMOD rep.),
- Marc Najork (Microsoft)
- Paolo Boldi (Univ. de Milano)
- Brian Davison (Lehigh Univ., SIGWEB rep.)
- Bing Liu (UIC, SIGKDD rep.)

January 2014: Hang Li (Huawei) joined to complete the SC and soon after I proposed
Marc Najork (Google) as new chair, who was elected by acclamation!

2017: SC should be partially renewed

WSDM’s History

<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Conference Chair</th>
<th>PC Chairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>Palo Alto</td>
<td>Marc Najork</td>
<td>Andrei Broder; Soumen Chakrabarti</td>
</tr>
<tr>
<td>2009</td>
<td>Barcelona</td>
<td>Ricardo Baeza-Yates</td>
<td>Paolo Boldi; Berthier Ribeiro-Neto</td>
</tr>
<tr>
<td>2010</td>
<td>New York</td>
<td>Brian Davison; Torsten Suel</td>
<td>Nick Craswell; Bing Liu</td>
</tr>
<tr>
<td>2011</td>
<td>Hong Kong</td>
<td>Irwin King</td>
<td>Wolfgang Nejdl; Hang Li</td>
</tr>
<tr>
<td>2012</td>
<td>Seattle</td>
<td>Eytan Adar; Jaime TEEvan</td>
<td>Eugene Agichtein; Yoelle Maarek</td>
</tr>
<tr>
<td>2013</td>
<td>Rome</td>
<td>Stefano Leonardi; Alessandro Panconesi</td>
<td>Paolo Ferragina; Aristides Gionis</td>
</tr>
<tr>
<td>2014</td>
<td>New York</td>
<td>Ben Carterette; Fernando Diaz</td>
<td>Carlos Castillo; Donald Metzler</td>
</tr>
<tr>
<td>2015</td>
<td>Shanghai</td>
<td>Xueqi Cheng; Hang Li</td>
<td>Evgeniy Gabrilovich; Jie Tang</td>
</tr>
<tr>
<td>2016</td>
<td>San Francisco</td>
<td>Paul Bennett; Vanja Josifovski</td>
<td>Jennifer Neville; Filip Radlinski</td>
</tr>
<tr>
<td>2017</td>
<td>Cambridge</td>
<td>Milad Shokouhi; Maarten de Rijke</td>
<td>Andrew Tomkins; Min Zhang</td>
</tr>
</tbody>
</table>
WSDM’s Growth: Papers

![Graph showing the growth of submitted and accepted papers from 2008 to 2017.](image1)

WSDM’s Quality: Acceptance Rate

![Graph showing the acceptance rate from 2008 to 2017.](image2)
WSDM’s Growth: People

WSDM’s Impact: ACM Citations & Downloads (Dec 2016)

100 to 380 downloads per year!
## WSDM’s Places

<table>
<thead>
<tr>
<th>Year</th>
<th>University Research</th>
<th>Microsoft</th>
<th>Yahoo</th>
<th>IBM</th>
<th>Laboratory</th>
<th>Indiana</th>
<th>Illinois</th>
<th>Technology</th>
<th>State</th>
<th>University Research</th>
<th>Microsoft</th>
<th>Institute</th>
<th>Google</th>
<th>UIUC</th>
<th>Technology</th>
<th>Yahoo</th>
<th>Tsinghua</th>
<th>CMU</th>
<th>UI at Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>University Research</td>
<td></td>
<td></td>
<td></td>
<td>Laboratories</td>
<td>Stanford</td>
<td>Technology</td>
<td>Corporation</td>
<td>Chinese</td>
<td>University Research</td>
<td>Technology</td>
<td>Microsoft</td>
<td>Science(s)</td>
<td>State</td>
<td>Yahoo</td>
<td>UIUC</td>
<td>Technology</td>
<td>Yahoo</td>
<td>Tsinghua</td>
</tr>
</tbody>
</table>

## WSDM’s Topics

| Year | Search | Web | Ranking | Mining | Classification | Advertising | Models | Document | Graph | Analysis | Search | Web | Social | Social Networks | Data | Modeling | Learning | Systems | Click | Advertising | Mining | Social Networks | Search | Learning | Time | IR | Modeling | Text | Data Mining | Recomm. Systems | Embedding |
|------|--------|-----|---------|--------|----------------|-------------|--------|----------|-------|---------|--------|-----|-------|---------------|------|-----------|----------|--------|------|------------|------|-------------|----------------|---------|
| 2008 |        |     |         |        |                |              |        |          |       |         |        |    |       |               |      |           |          |        |      |            |      |             |                |         |
| 2011 | Search | Web | Social  | Mining  | Query | Analysis | Temporal | Online | Learning | Data  |        | Search | Web |         | Social Networks | Data | Modeling | Learning | Systems | Click | Advertising | Mining | Social Networks | Search | Learning | Time | IR | Modeling | Text | Data Mining | Recomm. Systems | Embedding |
| 2014 |        |     |         |        |       |          |        |        |       |         |        |    |       |               |      |           |          |        |      |            |      |             |                |         |
| 2017 | Social Networks | Search |         |        | Data | Social Networks | Modeling | Learning | Systems | Click | Advertising | Mining |         | Social Networks | Search | Learning | Time | IR | Modeling | Text | Data Mining | Recomm. Systems | Embedding |
Current & Future Challenges

- Contextual Multilingual Semantic Search
- Fake web content
- Bias in the Web

Contextual Multilingual Semantic Search

- Multilingual Knowledge Base
- Contextual Query Understanding
- Semantic Ranking
Knowledge Base

Challenge: Multilingual
Only at Lexicon level

Query Understanding

Challenge: Multilingual
A NLP pipeline might be needed for semantic features
Semantic Ranking

Language dependence only on features

Fake Content & Bias

• British Prime Minister Benjamin Disraeli:
  • "There are three kinds of lies: lies, damned lies, and statistics.

UTC professor says "Everyone has bias"

Bias: significant deviation from a prior (unknown) distribution
(Observational) Human Data has Bias

- Gender
- Racial
- Sexual
- Religious
- Social
- Linguistic
- Geographic
- Political
- Educational
- Economic
- Technological

Goal: Bias Awareness

- from Noise or Spam
- Validity (e.g. temporal)
- Completeness
- Gathering process
- ...

Many people extrapolate results of a sample to the whole population (e.g., social media analysis)

In addition there is bias when measuring bias as well as bias towards measuring it!

Bias in the Web

Web

Data bias
Economic Bias in Links


Economic Bias in Links

[Baeza-Yates & Castillo, WWW2006]
### Website Structure

- **Brazil**: $\bar{x} = 3.5$, $\theta_1 = 0.8$, $\theta_2 = 2.9$
- **Chile**: $\bar{x} = 2.4$, $\theta_1 = 1.1$, $\theta_2 = 3.0$
- **Greece**: $\bar{x} = 3.3$, $\theta_1 = 1.1$, $\theta_2 = 2.9$
- **Korea**: $\bar{x} = 12.1$, $\theta_1 = 1.2$, $\theta_2 = 3.4$
- **Spain**: $\bar{x} = 0.1$, $\theta_1 = 1.5$, $\theta_2 = 2.5$

[Baeza-Yates, Castillo, Efthimiadis, TOIT 2007]

### Geographical Bias in Content

[E. Graells-Garrido and M. Lalmas, “Balancing diversity to counter-measure geographical centralization in microblogging platforms”, ACM Hypertext’14]
Gender Bias in Content

- Word embedding’s in w2vNEWS

Gender stereotype *she-he* analogies.

- sewing-carpentry
- nurse-surgeon
- blond-burly
- giggle-chuckle
- sassy-snappy
- volleyball-football

Gender appropriate *she-he* analogies.

- register-nurse-physician
- interior-designer-architect
- feminism-conservatism
- vocalist-guitarist
- diva-superstar
- cupcakes-pizzas

[Bolukbasi et al, ArXiv 2016]

Most journalists are men?

Yes, about 60 to 70% at work
although at college is the inverse

Gender Bias in Content

Systemic bias?

Equal opportunity?

Bias in the Web

Activity Bias

Which percentage of users produce 50% of the content?

<table>
<thead>
<tr>
<th>Platform</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>7%</td>
</tr>
<tr>
<td>Amazon Reviews</td>
<td>4%</td>
</tr>
<tr>
<td>Twitter</td>
<td>2%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]
Adding content implies adding wisdom?

We used Amazon’s reviews helpfulness and computed the text entropy

Content-based-wise users

How many of those users are being paid?

[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]
Content that is never seen: Digital Desert

- 1.1% of the Twitter content is never seen.*
- 31% of articles added/edited in May 2014 in Wikipedia, were not visited in June.

[Baeza-Yates & Saez-Trumper, ACM Hypertext 2015]
Sample Size?

- If we want to estimate the frequency of queries that appear with probability at least $p$ with a certain relative error $\xi$ we can use the standard binomial error formula $\sqrt{(1-p)/np}$ which works well for $p$ near $\frac{1}{2}$ but not for $p$ near 0.

- Better is the Agresti-Coull technique (also called take 2) which gives:

$$n \geq Z_{1-\alpha/2}^2 \left( \frac{p' (1-p')}{\epsilon^2} - 1 \right)$$

where $Z$ is the inverse of the standard normal distribution, $1 - \alpha$ is the confidence interval and $p' = p + Z^2 / 2$.

- If $p = 0.1$, $1 - \alpha$ is 90% and $\xi$ is 10%, we get $n = 2342$. The standard formula gives $n = 900$!

[Source: Baeza-Yates, SIGIR 2015, Industry track]

--

Sampling Techniques

- Standard technique: $p_q \approx \hat{p}_q(S) = \frac{f_q(S)}{\sum_{q' \in S} f_q(S)}$

- A good sample should cover well all the query distribution but this does not work with very biased distributions.

[Source: Zaragoza et al, CIKM 2010]
Incremental Stratified Sampling

- Main goal: make good samples consistent across time
- Simple idea based in stratified sampling: bins + random start point

![Bins](image)

- Bin size can be found by binary search starting with a good approximation if a query frequency model is used \((b < V/n)\)
- This perfectly mimics the head of the distribution, but not the tail
- Change the bins in the tail to get the right distribution

[Baeza-Yates, SIGIR 2015, Industry track]

---

Stratified Sampling Example

![Stratified Sampling Graph](image)
Extreme Algorithmic Bias

Bias in the Web

Data bias

Sampling bias

Algorithm

Algorithmic bias

Activity bias

Self selection bias

Interaction bias

Privacy
Bias in the Interaction

<table>
<thead>
<tr>
<th>Shop by Category</th>
<th>Position bias</th>
<th>Ranking bias</th>
<th>Presentation bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis Equipment</td>
<td></td>
<td></td>
<td>Sponsored</td>
</tr>
<tr>
<td>Tennis Games</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kids’ Sports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clothing, Shoes &amp; Jewelry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tennis - Books</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Bias in the Interaction**
  - Related Searches: tennis racket, tennis shoes.

- **Position bias**
  - **Ranking bias**
  - **Presentation bias**

- **Social bias**
  - **Interaction bias**
  - **Click bias**
  - **Scrolling bias**
  - **Mouse movement bias**
  - **Data & algorithmic bias**
  - **Self-selection bias**

Dependencies: A Cascade of Biases!

- **Position bias**
  - **Interaction bias**
  - **Click bias**
  - **Mouse movement bias**
  - **Scrolling bias**
  - **Data & algorithmic bias**
  - **Self-selection bias**
Social Bias

[WHY AMAZON’S RATINGS MIGHT MISLEAD YOU; The Story of Herding Effects
Ting Wang and Dashun Wang, Big Data, 2014]

Ranking Bias in Web Search

[Mediative Study, 2014]
Click Bias in Web Search

- Ranking & next page bias

### Unbiasing Search Clicks

Clicks as implicit positive user feedback

[Dupret & Piwowarski, SIGIR 2008]
[Chapelle & Zhang, WWW 2009]
Bias in the Web

Data bias

Activity bias

Sampling bias

Sparsity

Algorithmic bias

Second order bias

Interaction bias

(Self) selection bias

Privacy

Algorithm

Second Order Bias in Web Content

Ranking bias in new content
Redundancy grows (35%)

Person

Query

Search results

Web content is redundant (> 20%)

Clicks in results are biased to the ranking and the interaction

[Baeza-Yates, Pereira & Ziviani, Genealogical Trees in the Web, WWW 2008]
Avoid Second Order Bias due to Personalization

The Filter “Bubble”, Eli Pariser (2011)
- The effect of self selection bias
- Avoid the poor get poorer syndrome
- Avoid the echo chamber
- Empower the tail

Partial solutions:
- Diversity
- Novelty
- Serendipity
- Show me the dark side

Cold start problem solution: Explore & Exploit
How much exploration is needed for presentation bias?

Aggregating in the Tail

- Exploit the context (and deep learning!)
  91% accuracy to predict the next app you will use
  [Baeza-Yates et al, WSDM 2015]

- Personalization vs. Contextualization
  Recall that user interaction is another long tail

Tasks

Persons
It’s Hard to Get the Truth from Data (Professional Bias)

**Same Data, Different Conclusions**
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

- Statistically significant results showing referees are more likely to give red cards to dark-skinned players.
- One research team.
- 95% confidence interval.
- Referees three times as likely to give red cards to dark-skinned players.
- Twice as likely.
- Equally likely.
- Non-significant results.

61 analysts, 29 teams: 20 yes and 9 no (Univ. of Virginia, COS)

We need to focus on small data, not big data.

Questions?

**Biased Questions?**
More bias: We are hiring in Barcelona & San Diego!

Contact: rbaeza@acm.org
www.baeza.cl
@polarbearby