



KDD-2010

The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
WASHINGTON, DC • JULY 25-28, 2010

BioSnowball: Automated Population of Wikis

Xiaojiang Liu, **Zaiqing Nie**, Nenghai Yu, Ji-Rong Wen

Web Search and Mining Group
Microsoft Research Asia

Outline

- **EntityCube: Web-Scale Entity Summarization**
 - <http://entitycube.research.microsoft.com>
 - <http://renlifang.msra.cn> (Renlifang: Chinese version)
- **Related Work**
 - Wikipedia: Collaborative Editing
 - Decoupled: Fact Extraction and Biography Summarization
- **BioSnowball**
 - Bio-Fact Duality
 - Bootstrapping framework
 - Joint summarization
- **Experiments**
- **Conclusion**

Wikipedia: Collaborative Editing

- **Wikipedia**
 - Collaborative Editing
 - First stop for celebrities and notable entities
 - **Biography Introduction & Infobox**
 - **NPOV Policy**
 - Hard for everyday individuals

Bill Gates


From Wikipedia, the free encyclopedia

For other people named Bill Gates, see [Bill Gates \(disambiguation\)](#).
William Henry "Bill" Gates III (born October 28, 1955)^[2] is at the [world's wealthiest people](#)^[4] and was the wealthiest over remains the largest individual shareholder with more than £ Gates is one of the best-known entrepreneurs of the person which has in some cases been upheld by the courts (see [C](#) charitable organizations and scientific research programs t Bill Gates stepped down as chief executive officer of Microsoft full-time work at Microsoft to part-time work and full-time work. Gates' last full-time day at Microsoft was June 27, 2008. He

Contents [hide]

- 1 Early life
- 2 Microsoft
 - 2.1 BASIC
 - 2.2 IBM partnership
 - 2.3 Windows
 - 2.4 Management style
 - 2.5 Antitrust litigation
 - 2.6 Appearance in ads
- 3 Post-Microsoft
- 4 Personal life
 - 4.1 Philanthropy

Bill Gates



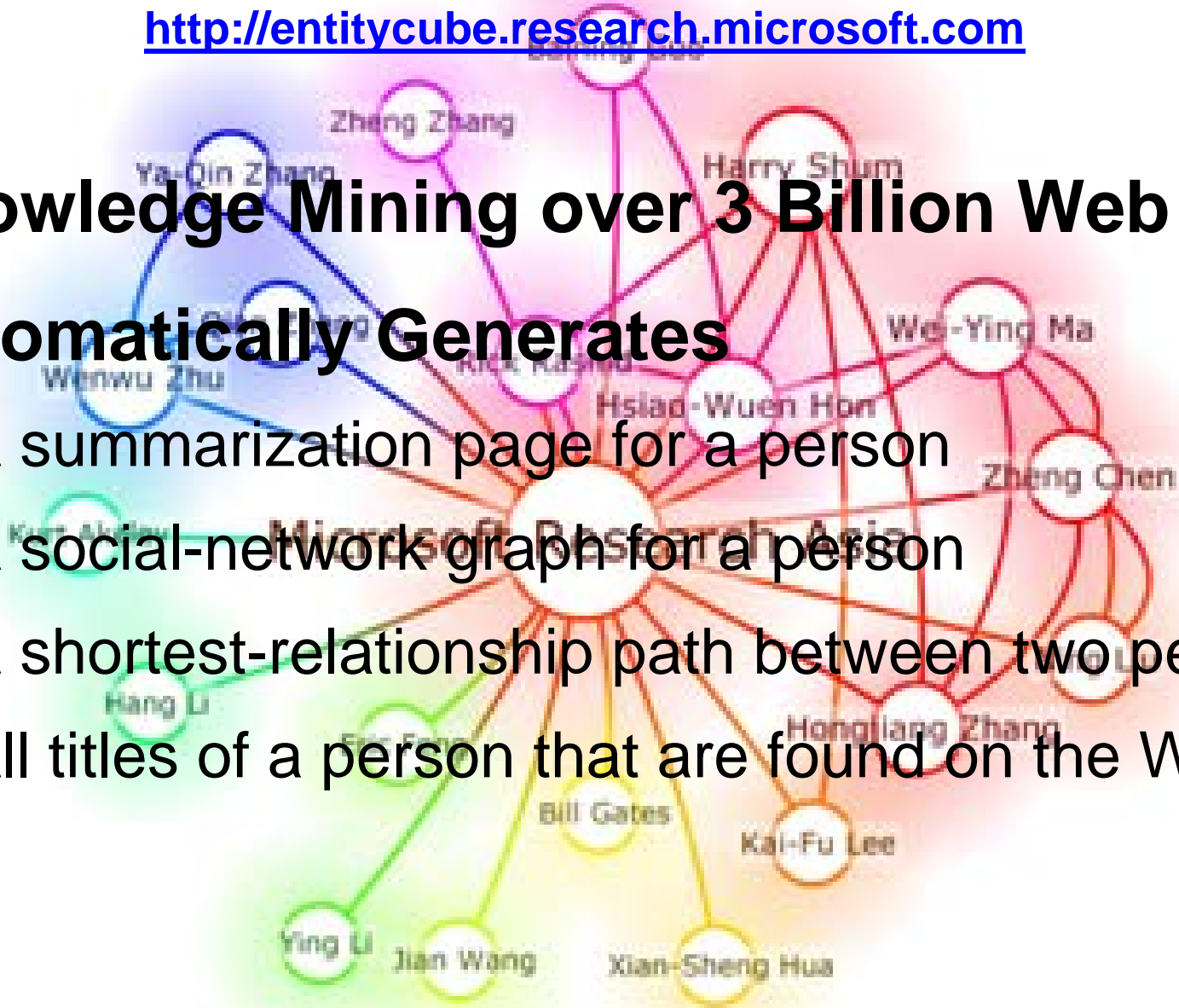
Bill Gates at the World Economic Forum in Davos, 2007

Born	October 28, 1955 (age 54) Seattle, Washington, USA
Residence	Medina, Washington, USA
Nationality	American
<i>Alma mater</i>	Harvard University (dropped out in 1975)
Occupation	Chairman of Microsoft (non-executive) Co-Chair of Bill & Melinda Gates Foundation Director of Berkshire Hathaway

EntityCube: Web-Scale Entity Summarization

<http://entitycube.research.microsoft.com>

- Knowledge Mining over 3 Billion Web pages
- Automatically Generates
 - ✓ A summarization page for a person
 - ✓ A social-network graph for a person
 - ✓ A shortest-relationship path between two people
 - ✓ All titles of a person that are found on the Web



Automated Fact and Biography Extraction

- **Separate Attempts**

- Fact Extraction

- StatSnowball: Jun Zhu, WWW'09
 - KnowItAll: Oren Etzioni, WWW'04

- Biography Summarization

- Sentence Extraction, L. Zhou, EMNLP, 2004
 - Information Extraction, M. Collins, NAACL-ANLP, 2000

- **De-coupled Attempts**

- Transitive and Latent Models for Biographic Fact Extraction, N. Garera, EACL'2009

- Multi-document summarization via information extraction, M. White, HLT'2001

Bio-Fact Duality

- **Good biography contains key facts**
- **If a text segment is a biography, fact extraction is easier**

William Henry “Bill” Gates III (born October 28, 1955) is an American business magnate, philanthropist, author, and chairman of Microsoft, the software company he founded with Paul Allen. He is ranked consistently one of the world’s wealthiest people and the wealthiest overall as of 2009. During his career at Microsoft, Gates held the positions of CEO and chief software architect, and remains the largest individual shareholder with more than 8 percent of the common stock. He has also authored or co-authored several books.

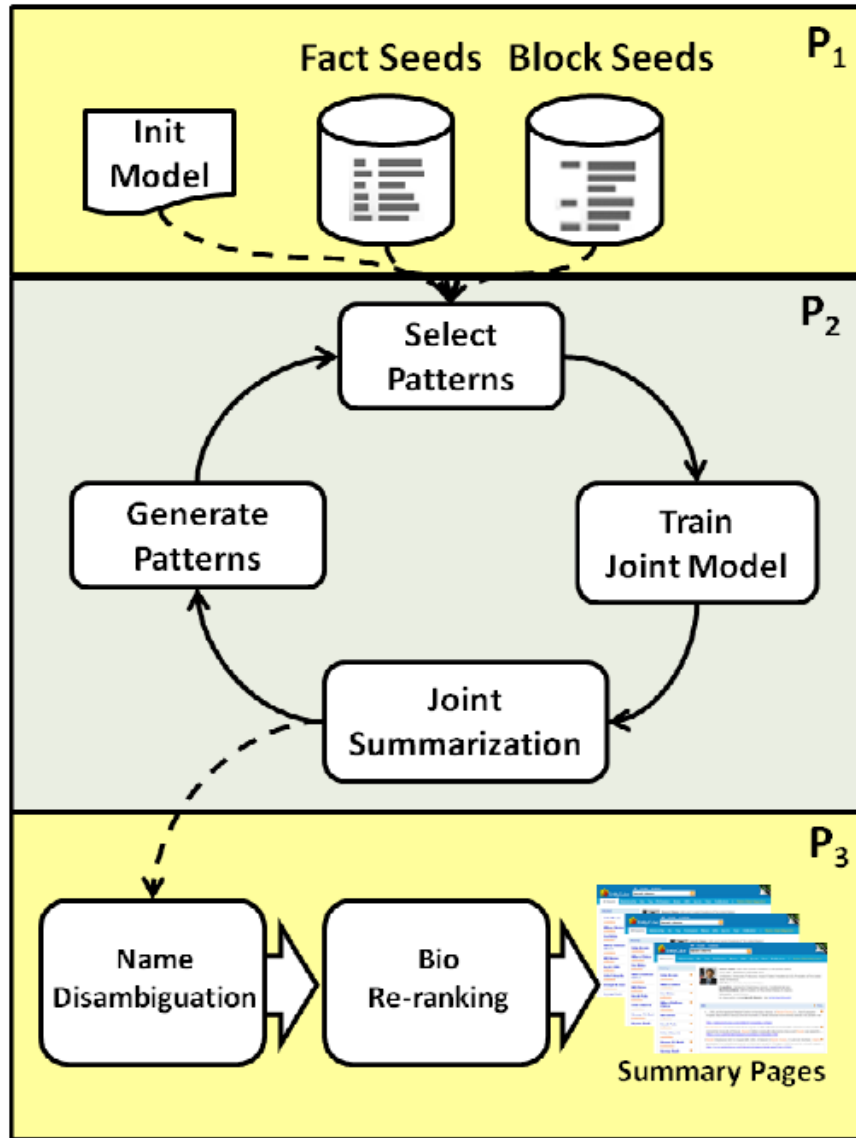
BioSnowball

- Jointly summarize facts and biographies
 - Bio-Fact Duality
 - Markov Logic Networks as the underlying model
 - probabilistic extension of first-order logic

$$p(\mathbf{q}|\mathbf{x}) = \frac{1}{Z(\mathbf{w}, \mathbf{x})} \exp \left(\sum_{i \in F_Q} \sum_{j \in G_i} w_i g_j(\mathbf{q}, \mathbf{x}) \right),$$

- discriminative model, easy to specify complex relations
- Bootstrapping Framework
 - Easy to get seeds
 - No need to get training data set

Architecture of BioSnowball



P₁ (Input):
initial fact and biography seeds
initial model can be provided

P₂ (Bootstrapping Summarization Model):
Four Bootstrapping steps
1. Select good patterns
2. Train a joint summarization model
3. Joint Summarize
4. Generate new patterns

P₃ (Post-processing and Output):
1. Use facts to do name disambiguation
2. Biography Re-ranking: remove duplicates

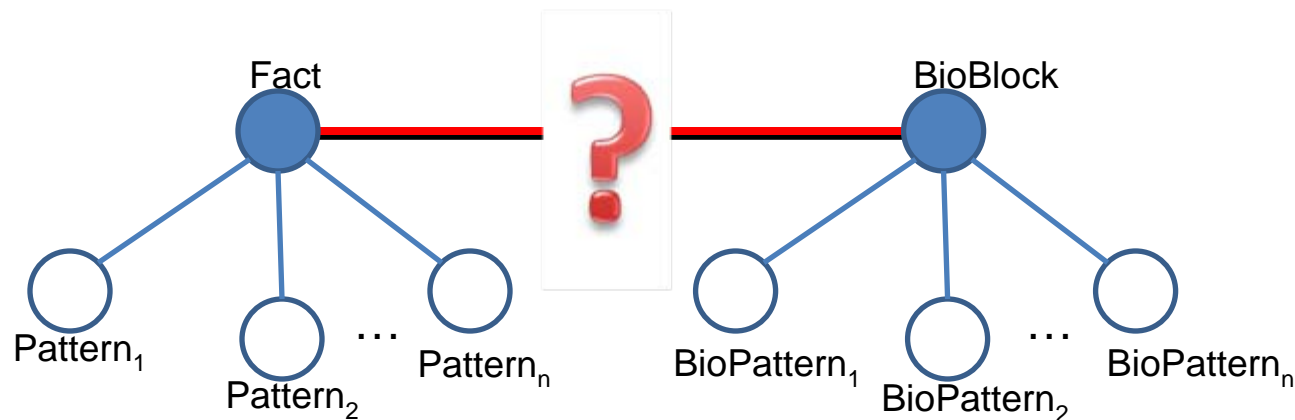
Joint Summarization Model

- **Fact Extraction**

- $Pattern(e, np, b, +r) \Rightarrow Fact(e, np, b, +r)$
- Can be general patterns or keywords patterns

- **Biography Ranking**

- $BioPattern(e, b) \Rightarrow BioBlock(e, b)$,
- Such as “is born”, “graduates from”



Joint Facts and Biographies

- **Fact can be used to Rank Biography**
 - $Fact(e, np, b, +r) \Rightarrow BioBlock(e, b)$
- **Co-reference in Biography**
 - $BioBlock(e, b) \wedge CoRef(e, pr) \wedge Pattern(pr, np, b, +r) \Rightarrow Fact(e, np, b, +r)$
- **Facts in Biography are in certain order**
 - $BioBlock(e, b) \wedge Fact(e, np, b, +r) \wedge Next(np, np) \Rightarrow Fact(e, np, b, +r)$

Experiments

- **Data set**
 - *WikiSeed*: 17850 with both infoboxes and bio-blocks from Wikipedia
 - *Web1M*: 1 million web blocks from EntityCube

Bio-Fact Duality

- **Most Frequent Fact Types**

Table 3: Top 10 Most Frequent Fact Types

Property	Occurrence	Hit in Bio	Hit Ratio
birthdate	16959	16529	0.975
name	15539	12557	0.808
birthplace	13762	13142	0.955
spouse	10888	5471	0.502
occupation	8022	6553	0.817
birthname	6897	5524	0.801
deathdate	6272	6128	0.977
deathplace	5319	4918	0.925
location	5013	3624	0.723
alma mater	3822	3145	0.823

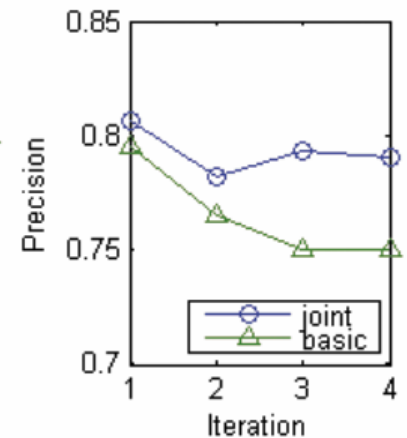
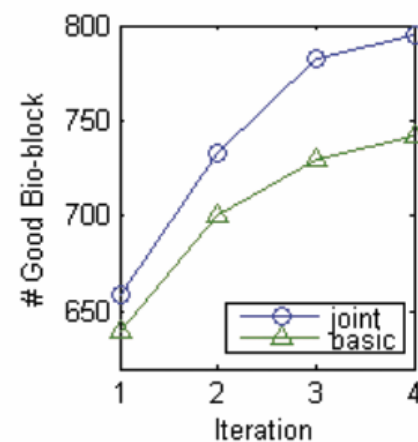
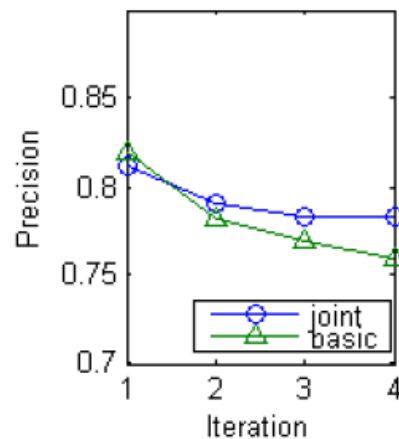
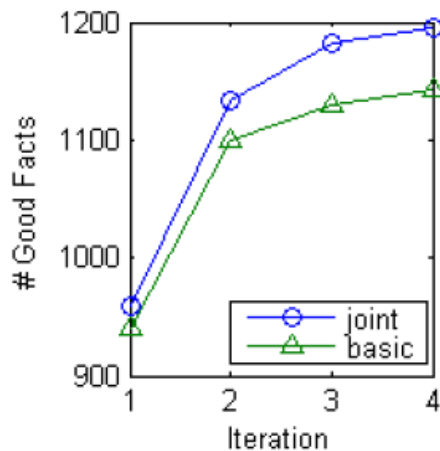
- **Biography Fact Position**

- First Block: 5.98 facts on average
- Second: 2.74 facts
- 3rd to 8th : 1 fact(s)

Joint Summarization & Bootstrapping Model

Table 4: Evaluation of fact extraction results of different joint summarization models

Types		Bio	Birth	Death	Overall Facts
bnBioSnowball	<i>Precision</i>	0.758	0.951	0.619	0.714
	<i>Recall</i>	0.675	0.791	0.325	0.754
	<i>F1</i>	0.714	0.864	0.426	0.734
jnBioSnowball	<i>Precision</i>	0.402	0.615	0.742	0.71
	<i>Recall</i>	0.380	0.064	0.442	0.08
	<i>F1</i>	0.390	0.116	0.554	0.144
jpBioSnowball	<i>Precision</i>	0.933	0.954	0.758	0.794
	<i>Recall</i>	0.770	0.932	0.543	0.908
	<i>F1</i>	0.844	0.943	0.633	0.847



Conclusions

- **EntityCube: Web-Scale Entity Summarization**
- **BioSnowBall**
 - Bio-Fact duality
 - Jointly perform biography ranking and fact extraction in an integrated statistical model.
 - A bootstrapping architecture
 - BioSnowball significantly reduces the number of human-tagged examples and iteratively mines facts and biography blocks.

Questions

Thank you!