Exploiting Statistical and Relational Information on the Web and in Social Media: Applications, Techniques, and New Frontiers

Part III: New Frontiers
Part III Roadmap

- Improving expressivity
  - PSL: Incorporating arbitrary aggregation functions into SRL models and reasoning about similarity

- Inference
  - Probabilistic databases
  - Lifted inference

- Learning
  - Online learning
  - Concept drift, emerging events
  - Privacy
Improving Expressivity

- Statistical relational languages are attractive:
  - Use a flexible relational language to define local and collective features
  - Provide a principled way for probabilistic reasoning

- However, in many Web apps it is also convenient to:
  - Easily incorporate arbitrary aggregation functions, e.g. the ones we defined in Part II
  - Reason not just about truth values but also about similarity

- One statistical relational model that also accommodates these desiderata is Probabilistic Similarity Logic (PSL)
Probabilistic Similarity Logic

Probabilistic reasoning in PSL

- Like other SR models, comes with an first-order logic language for expressing relational dependencies:

\[
\text{Category}(A, C) \iff \text{Category}(B, C) \land \text{Unknown}(A) \land \text{link}(A, B) \land A \neq B
\]

Reasoning about similarity

- Arbitrary similarity functions on entity attributes:

\[
A \approx B \iff A.\text{name} \approx B.\text{name}
\]

- Arbitrary similarity functions on relation-defined sets:

\[
A \approx B \iff \{A.\text{friends}\} \approx \{B.\text{friends}\}
\]

© Getoor & Mihalkova 2010
PSL: Some Details

- Similarities in a rule are combined using T-norms:
  - Lukasiewicz T-norm
    - $\oplus (h_1, h_2) = \min(1, h_1 + h_2)$
    - $\otimes (h_1, h_2) = \max(0, 1 - h_1 + h_2)$
  - can be customized

- Distance to satisfaction of a grounded rule
  - $d(R, I) = \max(\otimes (B_1, \ldots, B_n) - \oplus (H_1, \ldots, H_m), 0)$

- Distance to satisfaction of PSL program
  - Combine the distances of satisfaction of all rule instantiations

---

Slide credit: Adapted from slides by Matthias Bröcheler
PSL in Wikipedia

Graphic credit: Matthias Bröcheler
Wikipedia Rules

\[
\begin{align*}
\text{hasCat}(A,C) & \iff \text{hasCat}(B,C) \land A! = B \land \\
& \quad \text{unknown}(A) \land \text{document}(A,T) \land \\
& \quad \text{document}(B,U) \land \text{similarText}(T,U) \\
\text{hasCat}(A,C) & \iff \text{hasCat}(B,C) \land \text{unknown}(A) \land \\
& \quad \text{link}(A,B) \land A! = B \\
\text{hasCat}(D,C) & \iff \text{talk}(D,A) \land \text{talk}(E,A) \land \\
& \quad \text{hasCat}(E,C) \land \text{unknown}(D) \land A! = B
\end{align*}
\]

Slide credit: Matthias Bröcheler
Part III Roadmap

- Improving expressivity
  - PSL: Incorporating arbitrary aggregation functions into SRL models and reasoning about similarity

- Inference
  - Probabilistic databases
  - Lifted inference

- Learning
  - Online learning
  - Concept drift, emerging events
  - Privacy
Several approaches proposed in recent years in DB literature

- Annotate tuples with probabilities of existence (tuple-existence uncertainty)
- Specify a pdf over possible values of an attribute (attribute-value uncertainty)
- Focus on SQL query evaluation, but inference also considered
- Strong independence assumptions; limited attribute uncertainty support
Part III Roadmap

- Improving expressivity
  - PSL: Incorporating arbitrary aggregation functions into SRL models and reasoning about similarity

- Inference
  - Probabilistic databases
  - Lifted inference

- Learning
  - Online learning
  - Concept drift, emerging events
  - Privacy
Scaling learning to web-size data:
- One promising direction: online learning
- Many parameter estimation techniques can be extended for online learning
  - [Huynh & Mooney, StarAI10] (tomorrow) for a max-margin online learning technique applied to query disambiguation

Addressing concept drift and its variants in Web applications
- Detecting newly emerging events/topics
- Shift in intent due to news topics, holidays, other events
  - e.g. [Syed et al., NIPS09] propose disambiguating with regard to such events
SR information can also be exploited to understand breaches to user privacy on the Web.
Privacy breaches in networks

- Identity disclosure
  - A mapping from a record to a specific individual

- Attribute disclosure
  - Find attribute value that the user intended to stay private

- Social link disclosure
  - Participation in a sensitive relationship or communication

- Affiliation link disclosure
  - Participation in a group revealing a sensitive attribute value

Slide credit: Elena Zheleva
A public profile on Facebook

Basic Information
- Networks: The World Bank, Washington, DC
- Sex: Female
- Birthday: February 2
- Hometown: Washington, DC
- Political Views: Liberal
- Favorite Quotations: Normal people are people you don't know well.

Groups
- Member of: Bryn Mawr College Class of 1991, Dogs at the Astoria, The Trews, Sarah Palin is NOT Hillary Clinton, I have more Foreign Policy Experience than Sarah Palin, DC Foodies, Bryn Mawr College Alumna, PeaceCorpsConnect – Returned Peace Corps Volunteers, IDS Alumni: George Washington University, Thailand will always be the Kingdom of Thailand not the republic, International Finance Corporation / The World Bank Group, Peace Corps Thailand

Friends
- 78 friends
  - Julia Bucknall
  - Roi Weitz
  - David Pollak
Emily’s friends and groups

Group affiliations cannot be hidden!

[Zheleva, Getoor, WWW 2009]
Conclusion

- Web & Social Media inherently noisy and relational
- Described a set of well-suited tools for dealing with noisy, relational data
- However, as of yet, not many success stories

Enablers:
- Scaling
- Online Feature construction
- Dealing with dynamic data

Time is right: technology & data
- New platforms parallel processing
- More data
- Growing need for both personalization and privacy
Acknowledgements

- Thank you to the Search Group at Microsoft Research, Misha Bilenko, Matthias Bröcheler, Amol Deshpande, Nir Friedman, Ariel Fuxman, Aris Gionis, Anitha Kannan, Alex Ntoulas, Hossam Sharara, Marc Smith, Elena Zheleva and others for their slides, comments, and/or helpful discussions.

- Lily is supported by a CI Fellowship under an NSF CRA grant.

- The linqs group at UMD [http://www.cs.umd.edu/linqs](http://www.cs.umd.edu/linqs) is supported by ARO, KDD, NSF, MIPS, Microsoft Research, Google, Yahoo!

© Getoor & Mihalkova 2010