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

Lise Getoor & Lily Mihalkova
University of Maryland, College Park
Statistical Relational Learning and the Web

Challenges Addressed by SR Learning and Inference

- Multi-relational data
  - Entities can be of different types
  - Entities can participate in a variety of relationships
- Probabilistic reasoning under noise and/or uncertainty

Challenges Arising in Web Applications

- Entities of different types
  - E.g., users, URLs, queries
- Entities participate in variety of relations
  - E.g., click-on, search-for, link-to, is-refinement-of
- Noisy, sparse observations
Tutorial Goals

- Understand the interactions between SRL and Web/Social Media applications:
  - What are some sources of relational information on the Web?
  - How has such relational information been exploited in existing approaches?
  - To what extent are existing SRL techniques a good fit for the challenges arising on the Web?
  - What future developments would make these areas more closely integrated?
Tutorial Road Map

- Part I: Survey of statistical and relational info on the Web and in social media
  - E.g., various search applications, social networks, knowledge discovery applications
- Part II: Survey of learning and inference techniques that can handle such info
  - “Flat” relational approaches, collective classification, SRL models
- Part III: Future directions

© Getoor & Mihalkova 2010
Disclaimer

- *Not* an attempt to provide a complete survey of the Web, social media, or SRL literatures
  - 3.5 hours is not enough for this!

- We provide a *biased* view, motivated by our goal of identifying the interesting intersection points of SRL and Web/social media applications
Exploiting Statistical and Relational Information on the Web and in Social Media: Applications, Techniques, and New Frontiers

Part I: Survey of statistical and relational information on the Web and in social media
Part I Road Map

- Query logs & query log applications
  - Social networks/social media & applications
  - Social Networks + Query Logs
  - Knowledge Discovery on the Web
Query logs record the interactions of users with a search engine.

Typically weighted to indicate strength of relation.
Query Log Applications

- Query Relatedness
- Sessionization
- Keyword generation
- Clustering queries/query refinements by information need addressed
- Query Personalization/Disambiguation
- Click Models
- Many others:
  - e.g., predicting commercial/non-commercial intent, query advertisement matching [See Eugene Agichtein’s tutorial this afternoon, 😊]
Query Relatedness

Goal: Use query log data to infer semantic relatedness between queries

Intuitions:

- The actions users take after querying capture information about the implicit semantic relations between queries
- Two queries are similar if the sets of URLs clicked after searching for them are identical, overlapping, or one is subset of the other

[Baeza-Yates & Tiberi, KDD07]
Query Relatedness

[Baeza-Yates & Tiberi, KDD07]

Subset-URLs  Identical-URLs  Partial-Overlap-URLs (a.k.a. Share-URL)

© Getoor & Mihalkova 2010
Queries are represented as a bag-of-URLs. Edges are weighted by the cosine similarity between two queries.

[Copyright © Getoor & Mihalkova 2010]
Query Relatedness [Baeza-Yates & Tiberi, KDD07]

- Semantic relatedness is based directly on derived relations
- Confirmed the intuitions that
  - Identical-URLs > Subset-URLs > Share-URLs in terms of quality of inferred similarities
  - Relations between queries are more reliable if a larger number of clicks was observed for each query in the relation
Sessionization

- Two kinds of sessions:
  - Search session
    - Determined using time-outs
  - Logical session
    - The same search session may contain queries for more than one information-seeking intent or search mission
    - Logical sessions may:
      - straddle search sessions
      - be intertwined

- Goal: Use query logs to determine whether two queries are part of the same logical session or search mission
Sessionization

Features Derived From:
- Clicked-For
- Shares-Words
- Same-Session
- Precedes-In-Session
- Precedes-Temporally

Used to Learn to Predict:
- Precedes-In-Logical-Session
  [Boldi et al., CIKM08]
- Same-Logical-Session
  [Jones & Klinkner, CIKM08]
Sessionization: Features

- Relations are typically not used directly; rather features are defined over them.

[Boldi et al., CIKM08, Jones & Klinkner, CIKM08]

- Clicked-For
- Shares-Words
- Same-Session
- Precedes-In-Session
- Precedes-Temporally

Word/character similarity, such as:
- Number of common words/characters
- Cosine, Jaccard similarity
- Character edit distance
Sessionization: Features

- Relations are typically not used directly; rather features are defined over them.

For example:
- Number of sessions in which co-occur
- Variety of stats over co-occurrence sessions, e.g. average length, average position of queries
- Statistical test indicating significance of co-occurrence

© Getoor & Mihalkova 2010
Sessionization: Features

- Relations are not used directly; rather, features are defined over them.

Examples:
- Average time between queries
- Time between queries > threshold

© Getoor & Mihalkova 2010
Keyword Generation

- End goal: suggest a set of effective keywords for online advertising campaigns (called concepts)

- Identify a set of queries (the keywords) related to a particular concept

- Concept described as a small set of “seed” URLs that represent it
Keyword Generation

Intuition: a query is a relevant keyword to a concept if a random walk that starts from the query ends up at one of the concept’s URLs

[Fuxman et al., WWW08]

© Getoor & Mihalkova 2010
Clustering Query Refinements

- Query refinement
  - A re-writing of an initial query to capture meaning nuance more specifically
- Goal: Cluster possible refinements of a query by topic

[Sadikov et al., WWW10]
Clustering Query Refinements

- Intuition: Two queries cover the same topic if users tend to end up at the same URLs after searching for them
  - i.e., if a random walk through the graph ends up in the same URLs

[Sadikov et al., WWW10]
Modeling Information Need

- Query clusters can be viewed as sets of queries that fulfill the same information need.
- We can expand our relational graph to include an abstract object: the information need.
- Two kinds of relations with info need:
  - Query $\rightarrow$ Info need: Goal of the query is to fulfill info need.
  - URL $\rightarrow$ Info need: URL is relevant to particular info need.
- Info needs and relationships with them are unobserved.
  - However, main intuition of query log applications is that they are implicit in user actions.
Modeling Information Need

Clustering the queries/URLs can be viewed as inferring belongs-to-info-need relations.

Some queries are ambiguous and may be issued for more than one info need.

Some URLs may satisfy more than one need as well.
Recap with Info Need

- We can cast query log application discussed so far in terms of information need:
  - Query Relatedness: Find queries that serve similar information needs
  - Sessionization: Given a search session, identify the info need threads
  - Keyword generation: Generate keywords that satisfy particular information need
  - Query refinement clustering: Cluster query refinements by info need

© Getoor & Mihalkova 2010
Personalized Search

- Extend info need treatment to include users

© Getoor & Mihalkova 2010
Personalized Search

- Based on ternary relations between users, queries, and URLs

[Almeida & Almeida, WWW04]
[Sugiyama et al., WWW04]
[Dou et al., WWW07]
[Cao et al., WWW09]
[Mihalkova & Mooney, ECML09]

© Getoor & Mihalkova 2010
Collaborative Filtering Connection

Collaborative filtering

- Task: Recommend new items to users based on preferences of users with similar interests
- User similarity is inferred from commonalities in highly/lowly rated items
- e.g. [Breese, TechRep98; Herlocker, SIGIR99]

Personalized search

- Similarities:
  - Users → Users; Items → URLs
  - Clicking a URL is analogous to rating it highly

- Differences:
  - Clicks need to be considered in the context of queries
  - Sparser, noisier, harder to get reliable negative ratings

© Getoor & Mihalkova 2010
Click Models/Ranking

- The quintessential search engine problem
- Predict the probability that a URL is clicked, given that it is shown for a given query
- URLs with higher probability of being clicked/relevant ranked higher
- Lots of different approaches developed
  - Here we focus on some that use query log data to infer relevance
Click Models/Ranking

Precedes-In-Session
More-Relevant-Than
Seen-But-Skipped
Share-Terms
Re-define to include browsing behavior

[Joachims et al., SIGIR05]
[Radlinski & Joachims, KDD05]
[Agichtein et al., SIGIR06]
[Craswell & Szummer, SIGIR07]
[Bilenko & White, WWW08]
[Chapelle & Zhang, WWW09]
[Guo et al., WWW09]
...

© Getoor & Mihalkova 2010
Much research has focused on defining informative features on these relations:
- e.g., a small sample from [Agichtein et al., SIGIR06]

**Share-Terms**
- Term overlap between query and page title
- Term overlap between query and URL
- Term overlap between query and page summary

**Clicked-For**
- Position of clicked URL on the page of results
- Relative frequency of a click
- Is previous/next result clicked?
Click Models/Ranking

- The pattern of clicking or skipping a search result has been used to infer relevance of URLs.
- E.g., some examples from [Joachims et al. SIGIR05; Radlinski & Joachims, KDD05]
Summary of Query Logs Apps

© Getoor & Mihalkova 2010
Part I Road Map

- Query logs & query log applications
  - Social networks/social media & applications
- Social Networks + Query Logs
- Knowledge Discovery on the Web
Online Social Networks (OSNs)
Relational Info in OSNs

- Friends
- Collaborators
- Family
- Fan/Follower

Comments, Replies, Edits, Co-Edits, Co-Mentions, etc.

© Getoor & Mihalkova 2010
Tasks

- Finding important nodes, “influentials”
  - Key-opinion leader identification
- Understanding Behaviors
  - Viral marketing & Information Flow
  - Social Roles
  - Collaborative Dynamics, Social Computing
- Link Prediction
  - Link recommendation
- Community Discovery

© Getoor & Mihalkova 2010
Opinion Leaders

- A small number of individuals who influence an exceptional number of their peers*
- Play an important role in opinion formation


© Getoor & Mihalkova 2010
Tasks

- Finding important nodes, “influentials”
  - Key-opinion leader identification
- Understanding Behaviors
  - Viral marketing & Information Flow
  - Social Roles
  - Collaborative Dynamics, Social Computing
- Link Prediction
  - Link recommendation
- Community Discovery
Viral Marketing

© Getoor & Mihalkova 2010
Social Roles in Discussion Forums

- **Answer person**
  - Outward ties to local isolates
  - Relative absence of triangles
  - Few intense ties

- **Reply Magnet**
  - Ties from local isolates often inward only
  - Sparse, few triangles
  - Few intense ties

- **Discussion person**
  - Ties from local isolates often inward only
  - Dense, many triangles
  - Numerous intense ties

Slides courtesy of Marc Smith, Community Action
Social Roles in Yahoo Answers

Figure 4: Sampled ego networks of three selected categories

Variants of transitivity:

- If you are my friend, my enemies, are your enemies
- If we are enemies, your enemies are my friends

[Golbeck, 2005, iTrust06]: How does trust propagate in recommendation networks?

[Kunegis et al., WWW09]: Which are the unpopular users? What is the sign of a relationship between users?

[Guha et al., WWW04]: How does trust and distrust propagate?

[Leskovec et al., WWW10]: What are user attitudes (sign of relationship) toward one another?
Tasks

- Finding important nodes, “influentials”
  - Key-opinion leader identification
- Understanding Behaviors
  - Viral marketing & Information Flow
  - Social Roles
  - Collaborative Dynamics, Social Computing
- Link Prediction
  - Link recommendation
- Community Discovery

© Getoor & Mihalkova 2010
[Agrawal et al., WWW03]: Use the fact that this is typically an antagonistic relationship to infer separation into opposing camps.
nuanced collaborations

[Brandes et al., WWW09]: Studied editor interactions in Wikipedia
**Multimodal social networks**

**Online social network (OSN):**

- Bossa Nova

**Online affiliation network (OAN):**

- **Group 1:** Bossa Nova
- **Group 2:** I love Apple
- **Group 3:** Yucatan
Multimodal social networks

**Online social network (OSN):**

- One mode: user-user links

**Online affiliation network (OAN):**

- Two modes: user-group links

- Bossa nova
- I love 🍌
- Yucatan
Part I Road Map

- Query logs & query log applications
- Social networks/social media & applications
  - Social Networks + Query Logs
- Knowledge Discovery on the Web

© Getoor & Mihalkova 2010
Social Networks & Query Logs

Strength of relationship (amount of time spent talking) indicated by line thickness.

[Singla & Richardson WWW08]: Similarities between querying behavior and talking to each other or having friend in common.
Social Tagging

- Tag recommendation
  - Not personalized – tags are suggested regardless of who the user is
    - e.g., [Heymann et al., SIGIR08]
  - Personalized – tags are suggested to match each particular user’s interests
    - e.g., [Rendle et al., KDD09]

- Document/item recommendation using tags
  - Akin to collaborative filtering, but based on ternary relation between users, tags, and documents
    - e.g., [Sen et al., WWW09; Guan et al., WWW10, Shepitsen et al., RS08]
Social Tagging, View 1

- Ternary relationships between tags, users, documents
Social Tagging, View 2

- Tri-partite graph
  - Aggregate over documents/tags

[Shepitsen et al., RS08] [Guan et al., WWW10]

Document recommendations are based on not just preferences of similar users but also preferences for tags.
Part I Road Map

- Query logs & query log applications
- Social networks/social media & applications
- Social Networks + Query Logs
  - Knowledge Discovery on the Web
Knowledge Discovery on the Web

- Extracting & Discovering
  - Entities
  - Sets of entities
  - Relationships
  - Taxonomies
  - Knowledge bases
  - … bottom up construction of the Semantic Web
Folksonomy users commonly have in their mind (*hidden*).

Can we recover the folksonomy from many observed hierarchies?  \( \rightarrow \) folksonomy learning!

Personal hierarchies from various users (*observed*) such as users' folder-sub folders.

Users select a portion of the hierarchy to organize their content.

[shallow, noisy, sparse(incomplete) & inconsistent]

[Plangprasopchok et al. KDD10]
Refining Ontologies

- Use info gleaned from web sources to refine ontologies

Wikipedia Info Boxes, each representing a concept, described by various attributes

[Wu & Weld, WWW08]

© Getoor & Mihalkova 2010
Generalizing Concepts & Relations

- Extract relationship tuples from web text and organize them conceptually

[Kok & Domingos, ECML08]
Machine Reading Project

- Very Large-scale AI
- Extract common-sense knowledge from the web
- Involves \textit{self-supervised} extraction of \textit{entities, relations} from text

Work done at: UW, CMU, Stanford, ISI, UIUC, NYU, BBN, SRI, IBM, Cycorp, etc.

Slide courtesy of Oren Etzioni
Part I Summary

- Huge space of problems
  - Query logs & query log applications
  - Social networks/social media & applications
  - Social Networks + Query Logs
  - Knowledge Discovery on the Web

- All problems inherently relational, inherently noisy, large-scale, etc.