Social Content Exploration

Cong Yu

Yahoo! Research New York

Joint with Sihem Amer-Yahia

Panel on Social Network Mining & Search

ICDE, Shanghai, China

March 30, 2009

Three Messages

The Emergence of Social Content Sites

Recommendation is Emerging as One of the Dominant Information Exploration Paradigms

Developers Need Help Building Information Exploration Applications

Cong Yu
Yahoo! Research New York

Three Messages

The Rise of Social Content Sites

Cong Yu
Yahoo! Research New York

An Emerging Trend: Social Content Integration

Cong Yu
Yahoo! Research New York

Social Content Sites (SCS)

- Web destinations that let users:
  - Consume content-oriented information: videos, photos, news articles, etc.
  - Engage in social activities with their friends and people of similar interests
- Two major driving factors:
  - Incorporating social activities improves the attractiveness of traditional content sites
    - The “similar traveler” feature improves the user duration on Yahoo! Travel significantly
  - Incorporating content is critical to the value of social networking sites
    - A significant amount of user time is spent on browsing other people’s photos, notes, etc.

Cong Yu
Yahoo! Research New York

Implications

- Social information integration
  - Users are tired of having their online profiles replicated at many different places
  - A few Social Networking Sites will dominate and supply social information to most Social Content Sites
  - The challenge is not on how to integrate, but also on what to integrate!
    - Users do use same profile across sites (if they don't, then it's probably because they want to keep it separate)
    - Enormous amount of social activities, which ones to integrated for a particular Social Content Site?
- Privacy versus utility
  - What is the right granularity?
  - What is the right interface?

Cong Yu
Yahoo! Research New York

Three Messages

Recommendation is Emerging as One of the Dominant Information Exploration Paradigms

Cong Yu
Yahoo! Research New York

Search vs. Recommendation

	General (e.g., things to do)	Categorical (e.g., family)	Specific
with locations	32.36%	22.52%	8.37%
w/o locations	21.38%	5.34%	

Summary Statistics of 10 Million Yahoo! Travel Queries

- Majority of those searches are in fact *recommendations in disguise*!
  - Users are usually NOT looking for specific items
  - Majority of the sessions are ***seeking recommendations*** with ***geographical and topical constraints***

Cong Yu
Yahoo! Research New York

Information Exploration on SCS

- Major paradigms
  - **User-based:** browsing the content of your friends or other users you simply follow (Facebook, twitter)
  - **Search:** content based keywords matching (most traditional content sites)
  - **Recommendations:** hotlists, tag-based hotlists (Amazon, most collaborative tagging sites)
  - **Query:** database-style querying with complex conditions (not many so far)
- Ranking the paradigms:

  "Recommendation" ~ User-based > Search > Query

Cong Yu
Yahoo! Research New York

Implications and Our Work

- Topical/community analysis becomes a necessity
  - Keyword analysis is no longer enough
  - A good result will depend on:
    - The content itself and the current information needs
    - **Who the user is**
    - **Who the user is connected to**
    - **Who the user trusts or should trust**
- Relevance is no longer the only things that matters
  - Diversity
  - Out-of-the-Box
  - Serendipity
  - Explanation
- Our recent work
  - Community-based recommendation for Yahoo! Travel [internal]
  - Explanation based diversification [ICDE/EDBT]

Cong Yu
Yahoo! Research New York

Recommendation Diversification

- Joint with Laks Lakshmanan
- While relevance is important to recommendation, others are critical too:
  - *Novelty*: avoid returning results that users are likely to know already.
  - *Serendipity*: aim to return less relevant results that might give users a pleasant surprise.
  - ***Diversity*: avoid returning results that are too similar to each other.**
- Recommendation Diversification:

From the pool of candidate items, identify a list of items that are dissimilar to each other while maintaining a high cumulative relevance, i.e., *strike a good balance between relevance and diversity*.

Cong Yu
Yahoo! Research New York

Existing Solutions for Diversification

- **Attribute-Based Diversification**
  - Diversity semantics: pair-wise distance functions based on item attributes (e.g., movie attributes).
  - Combining with relevance:
    - Threshold either relevance or distance, maximize the other
    - Optimize an overall score as a weighted combination of relevance and distance
  - Algorithm
    - Perform traditional recommendation
    - Obtain the attributes of each candidate item and compute the pair-wise distance
    - Ad-hoc methods follow how diversity and relevance are combined

- **A Major Problem:**
  - Lack of attributes suitable for estimating distance between pairs of items: e.g., URLs in del.icio.us, photos on Flickr, videos on Vimeo.

Cong Yu
Yahoo! Research New York

Explanation-Based Diversification

- Intuition
  - Explanation is the set of objects because of which a particular item is recommended to the user.
  - Two items share similar explanations are likely to be similar to each other.
- Explanation for Item-Based Strategies

$$\texttt{Expl}(u,i) = \{i' \in \mathcal{I} \mid \textbf{ItemSim}(i,i') > 0 \;\&\; i' \in \textbf{Items}(u)\}$$

- Explanation for Collaborative Filtering Strategies (social!)

$$\texttt{Expl}(u,i) = \{u' \in \mathcal{U} \mid \textbf{UserSim}(u,u') > 0 \;\&\; i \in \textbf{Items}(u')\}$$

Cong Yu
Yahoo! Research New York

Explanation-Based Diversity

- Pair-wise diversity distance between two recommended items
  - Standard similarity measures like *Jaccard similarity* and *cosine similarity*
  - E.g. (Distance based on Jaccard similarity)

$$DD_u^J(i,i') = 1 - \frac{|\mathtt{Expl}(u,i) \cap \mathtt{Expl}(u,i')|}{|\mathtt{Expl}(u,i) \cup \mathtt{Expl}(u,i')|}.$$

- Diversity for the set of recommended results (S)

$$DD_u(S) = avg\{DD_u(i,i') \mid i,i' \in S\}$$

Cong Yu
Yahoo! Research New York

Benefits and Practicality of Explanation-Based Diversification

- Applicable to items without attributes or whose attributes are difficult to analyze
  - Common on social content sites
- Explanations are by-products of many recommendation processes
  - They can be maintained with little overhead

Result set similarities (vs attribute-based)

high group	Algorithm Swap	Algorithm Greedy
Jaccard Similarity	0.83	0.79
Kendall *tau* Similarity	0.77	0.71
medium group	Algorithm Swap	Algorithm Greedy
Jaccard Similarity	0.95	0.98
Kendall *tau* Similarity	0.91	0.95

See our poster at next session for details

Cong Yu
Yahoo! Research New York

Three Messages

Developers Need Help Building Information Exploration Applications

Cong Yu
Yahoo! Research New York

Building Recommendation Applications is Hard

- Scalability challenges
  - Analyzing a graph of 10's of thousands nodes is very different from analyzing 10's of millions nodes
  - Better to let the underlying system take care of that!
- Semantic challenges
  - Lots of trials and errors going on in search of a good formula/model
  - Need to have a faster way for development

- As database people, we know what that means!
  - Algebra
  - Declarative language

Cong Yu
Yahoo! Research New York

SocialScope: A Graph Based Logical Algebra Framework

- Joint with Laks Lakshmanan
- Designed for information discovery on social content sites
- Aim to provide a declarative way of specifying analysis and query tasks
  - Uniformity and flexibility
  - Opportunities for performance optimization
- Basic operators:
  - Node Selection (σ^N), Link Selection (σ^L)
  - Composition, Semi-Join
  - Node Aggregation, Link Aggregation
  - Details in [CIDR09]

Cong Yu
Yahoo! Research New York

A Simple Search Task

Find John's friends who have visited travel destinations near Denver and all their activities.

1. $G_1 = \sigma^L_{friend}(G \ltimes_{(\mathtt{src},-)} \sigma^N_{john}(G))$ — John's friends
2. $G_2 = \sigma^L_{visit}(G \ltimes_{(\mathtt{tgt},-)} \sigma^N_{C_3}(G))$, where $C_3 =$ (destination, 'near Denver'). — People visited Denver
3. $G_3 = G_1 \ltimes_{(\mathtt{tgt},\mathtt{src})} G_2$ // subset of John's friends — John's friends who visited Denver
4. $G_4 = \sigma^L_{activity}(G \ltimes_{(\mathtt{src},\mathtt{tgt})} G_3)$ — Their activities

Cong Yu
Yahoo! Research New York

Jelly:
A Language Over Social Content Sites

- Designed with a focus on community-centric information exploration applications
  - Most useful applications
- A restricted implementation of the SocialScope algebra
  - Based on nested relation model, instead of full graph model
- Built-in primitives for topic and community generation
  - Topic Generation, Community Extraction
  - Recommendation Generation
  - Group Generation, Explanation Generation

Cong Yu
Yahoo! Research New York

A Simple (Incomplete) Example

- **Topic Generation and Community Extraction**

generate topics **for** *item* **into** *topics*
from tagging R
using LDA (seed, th=0.8)
seed R.*item* group R.*tag* weight-with count()

generate communities **into** *experts*
from topics T, tagging R
where T.*.*item* = R.*item*
using jaccard-similarity (seed, th=0.7)
seed (R.*user,* T.*topic*) list R.*item*

- **Recommendation Generation**

generate recommendations **into** *candidates*
given user u, query q
from experts T
where Selected (T.*topic*, u, q)
using count-users (seed)
seed T.*.*item* list T.*user*

- **Information Presentation**

generate explanations **into** *results*
from candidates C, tagging T
where C.*item* = T.*item*
using identity (seed)
seed C.*item* list T.*user* weight-with count()

Cong Yu
Yahoo! Research New York

Three Messages

The Rise of Social Content Sites

Recommendation is Emerging as One of the Dominant Information Exploration Paradigms

Developers Need Help Building Information Exploration Applications

Cong Yu
Yahoo! Research New York