Extracting, Mining and Predicting Users’ Interests from Social Networks
Extracting, Mining & Predicting

Users’ Interests from Social Networks

Fattane Zarrinkalam, Hossein Fani, Ebrahim Bagheri
KDD, August 4, 2019, Anchorage, Alaska
Organizers

Fattane Zarrinkalam
Postdoctoral Fellow Laboratory of Systems, Software and Semantics (LS3), Ryerson University, Canada
@FattaneZ
fzarrinkalam@ryerson.ca
http://ls3.rnet.ryerson.ca/people/fattane_zarrinkalam/

Hossein Fani
Ph.D. Candidate at University of New Brunswick
@HosseinFani
hfani@unb.ca
https://sites.google.com/site/hosseinfani/

Ebrahim Bagheri
Associate Professor & Director for the Laboratory for Systems, Software and Semantics (LS3)
Ryerson University, Canada
Canada Research Chair in Software and Semantic Computing
NSERC/WL Industrial Research Chair in Social Media Analytics
@ebrahim_bagheri
bagheri@ryerson.ca
http://ls3.rnet.ryerson.ca/
Our lab has been very active in industrial collaborations
  • Social Analytics
  • Web Mining
  • Semantic Technologies
  • Analysis of User Generated Content

We have worked with over 15 industrial partners
  • Over $9 Million project value
  • 65 trained HQP

Our lab, LS$^3$, currently has 30 members
We are proud of our over 65 alumni
Outline

1. Introduction

2. User Interest Modeling

3. Challenges and Future Directions
Outline

1 Introduction

2 User Interest Modeling

3 Challenges and Future Directions
Motivation

Many service providers are now focused on *customized* and targeted *personalization* of content for their end-users.

- Job recommendation in LinkedIn
- Friend recommendation in Facebook
- “Who to Follow” service in Twitter

The main step in the personalization and recommender systems is *user interest modeling*
User Interest Modeling: Definition

- **User Interest Profile**: a data structure that represents interest characterization of a user at a particular moment in time.

- **User Interest Model**: definition and rules for the interpretation of the observations about the user & about the translation of the interpretation into the characteristics in a user interest profile.

- **User Interest Modeling**: the process of creating user interest profile following the definition and rules of the user interest model.
User Interest Modeling via Description

TED Recommends
Talks recommended just for you, delivered to your inbox

Now, tell us what you’re looking for

- Insights about issues that matter
- Fresh insights and knowledge
- Smart entertainment
- A sense of hope
- Professional growth
- A glimpse into the future
- A new perspective
- Inspiration or motivation
- Ideas for self-improvement

2/2
Already have a TED account? Sign in to see your recommendations
User Interest Modeling via Description

Collect facts about a user described by herself

- Take a lot of time to accumulate all the knowledge
- User may not always be able to provide accurate answers
  a) Doesn't know!
  b) Doesn't want to talk about it!
    - e.g., a student is interested in homosexuality
- Lacks the ability to automatically adapt to shifts in users’ interests

User interest modeling should not rely heavily on answers explicitly given by the user! It should infer user interest profile based on their activities, and traces on the web.
User Interest Modeling via Online Social Networks

1. Facebook
   - Monthly Active Users: 2.4 Billion
   - Daily Active Users: 1.6 Billion
   - Mobile Users: 88%
   - Daily Time Spent: 58min
   - Est. 2004

2. YouTube
   - Monthly Active Users: 1.9 Billion
   - Daily Active Users: 149 Million
   - Video Views Daily: 5 Billion
   - Average Visit Duration: 40 minutes
   - Est. 2005

3. WhatsApp
   - Monthly Active Users: 1.5 Billion
   - Daily Active Users: 1 Billion
   - New Users: 1 Million
   - Messages Daily: 60 Billion
   - Est. 2009

4. Instagram
   - Monthly Active Users: 1 Billion
   - Daily Active Users: 600 Million
   - Stories Active Users: 500 Million
   - Posts Daily: 95 Million
   - Est. 2010

5. Twitter
   - Monthly Active Users: 330 Million
   - Daily Active Users: 134 Million
   - Daily New Accounts: 460,000
   - Daily Tweets: 140 Million
   - Est. 2006

6. Reddit
   - Monthly Active Users: 330 Million
   - Views Per Month: 14 Billion
   - Active Communities: 138,000
   - Votes Daily: 25 Million
   - Est. 2005

7. LinkedIn
   - Monthly Active Users: 303 Million
   - Monthly New Accounts: 5.3 Million
   - Company Pages: 30 Million
   - Average Visit Duration: 10 minutes
   - Est. 2002

8. Snapchat
   - Monthly Active Users: 301 Million
   - Daily Active Users: 109 Million
   - Daily Video Views: 10 Billion
   - Daily Snaps: 3 Billion
   - Est. 2011

9. Pinterest
   - Monthly Active Users: 291 Million
   - Total Boards Created: 1 Billion
   - Total Pins Created: 175 Billion
   - Average Visit Duration: 14 minutes
   - Est. 2010

https://dustinstout.com/social-media-statistics/
Tutorial Scope

Don'ts

Photo | Video Social Media Platforms
Topic modeling approaches
User’s expertise detection
User embedding methods
  • low interpretable user representation

This tutorial is not an exhaustive account of works in related areas.
Outline

1 Introduction

2 User Interest Modeling

3 Challenges and Future Directions
2 User Interest Modeling

• Information Sources
• User Interest Representation Units
• User Interest Modeling Types
• User Interest Datasets
• Evaluation Methodologies
2 User Interest Modeling

- Information Sources
- User Interest Representation Units
- User Interest Modeling Types
- User Interest Datasets
- Evaluation Methodologies
Information Source

- Internal
- External
Information Source

- Internal
- External
Internal Information Source

- **Single-OSN**
  - LinkedIn
  - Facebook
  - Google+
  - Twitter
    - Widely used in most studies
      - Popularity
      - High degree of openness
    - ...

- **Multi-OSNs**
  - Cross-system user interest modeling
Fattane Zarrinkalam
@FattaneZ  Follows you

PostDoc researcher at LS3 laboratory. Interested in Information Retrieval, Social Media Mining, Semantic Web

Toronto, Ontario

Joined January 2015

Born September 02

Tweets

139

Following

120

Followers

148

Likes

816

Lists

2

Tweets

Fattane Zarrinkalam Retweeted

ACM SIGIR 2019 @sigir2019 · Apr 29

Check SIGIR 2019 tutorials on our website: sigir.org/sigir2019/prog...

Media

Fattane Zarrinkalam @FattaneZ · Apr 28

How to keep human bias out of AI
Fattane Zarrinkalam Retweeted

ACM SIGIR 2019 @sigir2019 · Apr 29
Check SIGIR 2019 tutorials on our website: sigir.org/sigir2019/prog...

Fattane Zarrinkalam @FattaneZ · Apr 28
How to keep human bias out of AI
Fattane Zarrinkalam
@FattaneZ  Follows you

PostDoc researcher at LS3 laboratory. Interested in Information Retrieval, Social Media Mining, Semantic Web

📍 Toronto, Ontario
🔗 ls3.rnet.ryerson.ca/people/fattane...
📅 Joined January 2015
📅 Born September 02
<table>
<thead>
<tr>
<th>Subscribed to</th>
<th>Member of</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>InformationRetrieval</strong> by Svitlana Vakulenko</td>
<td></td>
</tr>
<tr>
<td>522 Members</td>
<td></td>
</tr>
<tr>
<td><strong>ECIR</strong> by Svitlana Vakulenko</td>
<td></td>
</tr>
<tr>
<td>94 Members</td>
<td></td>
</tr>
</tbody>
</table>
Users use different OSNs for different purposes:

- Users on Facebook connect to their friends
- LinkedIn users connect to their business partners
- Twitter users might tweet about both recent news and events
- Tag-based Flickr profiles are related to geographical topics
- ...
Comprehensive understanding of user interests

- The overlap of a user's interests from different OSNs is very small so that the different profiles of a user complement each other

Solve cold-start and data sparsity problems in predictive tasks

- Using the information of Google+ to improve recommendations in Twitter (Piao and Breslin, UMAP’16)

Require user identity linkage across OSNs (Shu et al., KDD’17)

- Profile Inconsistency
- Content Heterogeneity
- Network Diversity
Internal Information Source: Multi-OSNs

Abel et al. Cross-system user modeling and personalization on the social web. UMUAI 2013.
Internal Information Source: Multi-OSNs

Abel et al. Cross-system user modeling and personalization on the social web. UMUAI 2013.
Internal Information Source: Multi-OSNs

Information Source

- Internal
- External
Analyzing social posts is challenging since they are short, noisy and informal.

• Content enrichment with external data
  • Related news articles (Abel et al., ESWC’11)
    • More than 85% of the tweets in the Twitter network are related to news
  • The content of embedded links/URLs in a post (Piao and Breslin, EKAW’16)
  • Knowledge bases such as Wikipedia (Kapanipathi et al., ESWC’14)
Check SIGIR 2019 tutorials on our website:
sigir.org/sigir2019/prog ...
Check SIGIR 2019 tutorials on our website:
sigir.org/sigir2019/prog...
User Interest Modeling

• Information Sources
• User Interest Representation Units
• User Interest Modeling Types
• User Interest Datasets
• Evaluation Methodologies
User Interest Representation Unit

- Keyword
- Group of Keywords
- Concept
- Group of Concepts
User Interest Representation Unit

► Keyword
► Group of Keywords
► Concept
► Group of Concepts
Each topic of interest is represented by a keyword (unigram | #tags) mentioned in the textual content

- Simple & predominant approach
- Suffers from the curse of dimensionality
- Sparse representation
- Forgo the underlying semantics of textual content
- Suffers from Polysemy and Synonymy
SIGIR

- Special Interest Group on Information Retrieval
- Special Inspector General for Iraq Reconstruction
User Interest Representation Unit

- Keyword
- Group of Keywords
- Concept
- Group of Concepts
Topic of interest is a probability distribution over keywords

- **Latent Dirichlet Allocation (LDA)**
  - **User-LDA**
    - All the posts from each user as a single document
  - **Post-LDA**
    - Each post as a separate document
    - All the post-topic distribution vectors by the same user are aggregated (e.g., by averaging)

- Post-LDA often learns better user representations than User-LDA in downstream applications (Ding et al., EMNLP’17)
Check SIGIR 2019 tutorials on our website: sigir.org/sigir2019/prog ...

5:10 AM - 29 Apr 2019

From @TheCable_FP Did the U.S. government buy favorable coverage of Iraq’s Anbar Province? http://bit.ly/4p6oj4 (from SIGIR report)

4:10 PM - 4 Nov 2009
Overcomes polysemy and synonymy limitations

Forgo the underlying semantics of textual content

Topic modeling approaches may not perform so well on posts
  • Designed for regular documents
  • Implicitly use co-occurrence patterns

Suffer from sparsity problem
User Interest Representation Unit

- Keyword
- Group of Keywords
- Concept
- Group of Concepts
Topic of interest is represented as a concept in a Knowledge Base

- Entity
- Category
- Hybrid
Representation via Concept: Entity

Using existing entity linking tools to extract entities from information sources such as user's tweets.
Representation via Concept: Entity

Check SIGIR 2019 tutorials on our website: sigir.org/sigir2019/prog ...

5:10 AM - 29 Apr 2019
16 Retweets 15 Likes

Special Interest Group on Information Retrieval

From Wikipedia, the free encyclopedia

SIGIR is the Association for Computing Machinery's Special Interest Group on Information Retrieval. The scope of the group's specialty is the theory and application of computers to the acquisition, organization, storage, retrieval and distribution of information; emphasis is placed on working with non-numerical information, ranging from natural language to highly structured data bases.

Contents
1 Conferences
  1.1 SIGIR conference locations
2 Awards
3 See also
4 References
5 External links

Conferences

The annual international SIGIR conference, which began in 1978, is considered the most important in the field of information retrieval. SIGIR also sponsors the annual Joint Conference on Digital Libraries (JCDL) in association with SIGWEB, the Conference on Information and Knowledge Management (CIKM), and the International Conference on Web Search and Data Mining (WSDM) in association with SIGKDD, SIGMOD.

Special Inspector General for Iraq Reconstruction

From Wikipedia, the free encyclopedia

The Office of the Special Inspector General for Iraq Reconstruction (SIGIR) (October 2004 – October 2013) was created as the successor to the Coalition Provisional Authority Office of Inspector General (CPA-IG). SIGIR was an independent government agency created by the Congress to provide oversight of the use (or misuse) of the $82 billion U.S. reconstruction program in Iraq. Stuart W.Bowen, Jr. was appointed to the position of CPA-IG on January 30, 2004 and served until its closure in October 2013. SIGIR reported directly to Congress, the Secretary of State, and the Secretary of Defense.

SIGIR's mission was to provide independent and objective oversight of U.S.-funded Iraq reconstruction policies, programs, and operations through comprehensive audits, inspections, and investigations. As of July 2006, SIGIR has issued 20 Quarterly Reports to Congress, 302 audits and inspections, 368 recommendations, and four Lessons Learned reports. SIGIR representatives have also testified before Congress on 27 separate occasions. Moreover, SIGIR's investigative and oversight work has resulted in 29 criminal indictments, more than $81 million in U.S. taxpayer funds saved or recovered, and $224 million being put to better use.

In February 2009, SIGIR issued its fourth Lessons Learned report, Hard Lessons: The Iraq Reconstruction Experience. Hard Lessons provides the first comprehensive account of the U.S. reconstruction effort in Iraq, chronicling the myriad challenges that confronted the rebuilding program, and concludes with 13 lessons drawn from the reconstruction experience.

Influence on Law and Policy. SIGIR reports have led to several important changes in U.S. reconstruction policy. These changes to the law and to key agencies' policies and procedures have increased management
Representation via Concept: Category

Represents more general user interests compared to using entities

► Wikipedia | DBpedia Categories
  • The relation between entities and categories are explicitly presented in the Wikipedia
  • Categories do not keep up with the real time nature of social networks (e.g. Twitter)

► News Categories
  Social networks and news media are similar in that many current issues are posted in both

► Open Directory Project (DMOZ) Categories
  Categories in this taxonomy provide a clear, meaningful and broad coverage of various real-world interests
Instead of using a single interest unit (entities xor categories), hybrid approaches combine different interest units

- Dbpedia entities and categories (Faralli et al., J. Web Semantics 2017)
- Dbpedia entities and WordNet synsets (Piao and Breslin, CIKM’16)
Background knowledge of these concepts can be exploited

- To extend the user interests by considering the relationship between concepts
- To characterize the user interests, e.g. measure the specificity of user interests based on links in DBpedia (Orlandi et al., WI’13)

Interests are confined to a set of predefined concepts

E.g. On July 22, 2019, W. Bruce Croft will deliver a talk on “The Importance of Interaction for Information Retrieval” in SIGIR 2019. There is no single concept in Wikipedia for this event.
Representation via Group of Concepts

Each topic of interest is represented by a group of concepts.
User Interest Modeling

- Information Sources
- User Interest Representation Units
- User Interest Modeling Types
- User Interest Datasets
- Evaluation Methodologies
User Interest Modeling Type

- Explicit
- Implicit
- Future
User Interest Modeling Type

- Explicit
- Implicit
- Future
“You are what you share.
— Charles Leadbeater
Explicit User Interest Modeling

Leveraging information from the user's own activities

► Textual contents of the user
  • E.g., a user mentions SIGIR frequently in her tweets

► User relationships
  • E.g., a user follows the Twitter account @sigir2019
Explicit User Interest Modeling ... Sample I


(a) Generic solution

(b) Example scenario
Explicit User Interest Modeling … Sample I


URL-based

Content-based

(a) Generic solution

(b) Example scenario
Explicit User Interest Modeling … Sample I

Explicit User Interest Modeling ... Sample I

Explicit User Interest Modeling … Sample II


Documents in news categories
(E.g. sports, food, politics, IT etc.)

News Corpus

Term-based feature generator

Terms
(E.g. Pizza)

Term vectors for news categories

Semantic gap

Wikipedia-based feature generators

Term vectors for messages

Wikipedia vectors for news category

Wiki-cluster

Wiki-category
(E.g. [ Category : Food (0.2),
Category : World cuisine (0.1), … ])

Wiki-article
(E.g. [ Article : Pasta (0.1),
Article : Omelette (0.05), … ])

Clustering

Similarity

User interest vector

(E.g. [ Food (0.8),
Sport (0.1),
IT (0.05), … ])
Explicit User Interest Modeling … Sample II

Explicit User Interest Modeling … Sample II

Explicit User Interest Modeling … Sample II

User Interest Modeling Type

- Explicit
- Implicit
- Future
Implicit User Interest Modeling

[explicit modeling]

“... learn what people say they like, not what people actually like!”

— Cuil’s CEO, Tom Costello
Implicit User Interest Modeling

Users’ implicit interests are those potential interests that the user did not explicitly mention but might have interest in.

- Not only improve user interest modeling for active users, but also for:
  - Free-riders (passive users): inactive generator, but active consumer
  - Cold start users: newly joined users

- Specially when a significant portion of social network’s users are passive
  - 4 in 10 users browse Facebook only passively, without posting anything
  - 44% of Twitter users have never sent a tweet
Implicit User Interest Modeling

- **Inter-user Relation**
  - “You are who you follow.
    - Homophily principle: users tend to connect to users with common interests
  - “People who are similar to me can reveal information about me.”

- **Inter-topic Relation**
  - User’s topics of interest are *semantically* related to each other
Implicit User Interests by Inter-user Relation

Inter-user Relation in Twitter:

- Follow
- Retweets
- Mentions

Retweeting behavior of a user is a stronger indicator of her topical interest compared to her following behavior (Welch et al. WSDM’11)
Implicit User Interests by Inter-user Relation

Using the user’s friends information

• Social posts (Wang et al., TIST 2014)

• Biography (Piao and Breslin, ECIR’17)
  - E.g. a user might be interested in *Information Retrieval* if she is following who describes herself as an *Information Retrieval researcher* in her biography in Twitter

• List membership (Piao and Breslin, HT’17)
  - E.g. a user might be interested in *Information Retrieval* if she is following who have been added into many topical lists related to *Information Retrieval*
Using predefined relation between concepts in Knowledge bases

Hierarchical (Kapanipathi et al., ESWC’14)
- E.g. Wikipedia category hierarchy
  - Represent user interests at different levels of granularity
  - Needs converting Wikipedia category structure to hierarchy

Graph-based (Piao and Breslin, HT ’17)
- E.g. Linked Open Data (LOD)
  - Provide more strategies based on different types of relations
  - The level of granularity of interests is not specified
Implicit User Interests by Inter-topic Relation

Measure collaborative relatedness between users’ topics of interest by considering users’ overlapping contributions toward topics

- **Collaborative Filtering** (Zarrinkalam et al., ECIR’16)
- **Frequent Pattern Mining** (Trikha et al. ECIR’18)
Identify user’s interests for non-famous users based on the expertise of their famous friends

- Non-famous users: less than 2K followers
- Famous users: at least 2K followers

1. Find the topical expertise of famous Twitter users

2. Extract interest tags for non-famous users using a *supervised* LDA model, Bi-Labeled LDA
1. Find the topical expertise of famous Twitter users (Bhattacharya et al., RecSys ’14)
   i. Collect all the Lists which have the user as a member
   ii. Calculate frequently occurring terms (unigrams and bigrams) from the List names and descriptions.
   iii. If a term has frequency more than 10, the user is identified as an expert on the term (topic)
   iv. The term is regarded as a tag of user.

E.g., Barack Obama is famous user and an expert in topics such as ‘politics’, ‘government’, ‘celeb’, ‘leader’
2. Extract interest for non-famous users using a \textit{supervised} LDA model, Bi-Labeled LDA

- A famous user may be expert at or famous in several aspects
- A non-famous user may follow a famous user due to only one aspect

E.g., Lance Armstrong is famous as a ‘world-class cyclist’ and a ‘cancer survivor’. A follower may be interested in:

- Cycling | Charity | Cycling & charity with different weights
Implicit User Interest Modeling … Sample I

He et al., Extracting Interest Tags for Non-famous Users in Social Network, CIKM’15.

Bi-labeled LDA
Implicit User Interest Modeling ... Sample I

- Document = Non-famous user
- Word happens in document = Following famous user
- $N_u = \text{Followings of user } u \text{ who are famous user}$

He et al., Extracting Interest Tags for Non-famous Users in Social Network, CIKM’15.

LDA (Blei et al., JMLR 2003)
• Non-famous user topics (tags) correspond to its famous user following’s tags

Labeled-LDA (Ramage et al. EMNLP’09)
• Word distribution of a topic is restricted to be only over those famous users who have this topic.
Implicit User Interest Modeling ... Sample II


1. Map a user’s followees to Wikipedia entities
2. Connect Wikipedia entities to Wikipedia category graph
3. Graph pruning to remove cycles

Interest profile
Implicit interests
Explicit interests
User’s followees
Implicit User Interest Modeling … Sample II

1. Map a user’s followees to Wikipedia entities

2. Connect Wikipedia entities to Wikipedia category graph

3. Graph pruning to remove cycles

1. **Selection of candidate senses**
   Finding a (possibly empty) list of candidate wikipages, using BabelNet synonym sets

2. **BoW Disambiguation**
   Similarity between the user biography and each candidate

3. **Structural Similarity**
   Similarity between the user’s friendships mapped to entities and candidates
Implicit User Interest Modeling ... Sample II


1. Map a user’s followees to Wikipedia entities
2. Connect Wikipedia entities to Wikipedia category graph
3. Graph pruning to remove cycles

User’s followees

Interest profile

Implicit interests

Explicit interests

Implicit User Interest Modeling...

Sample II

Implicit User Interest Modeling ... Sample II


(1) Map a user’s followees to Wikipedia entities

Interest profile

Graph pruning to remove cycles

Implicit interests

(2) Connect Wikipedia entities to Wikipedia category graph

Explicit interests

(3) User’s followees

Graph pruning to remove cycles
Implicit User Interest Modeling … Sample II

1. Map a user’s followees to Wikipedia entities
2. Connect Wikipedia entities to Wikipedia category graph
3. Graph pruning to remove cycles

User’s followees

Implicit interests

Explicit interests

Interest profile

Implicit User Interest Modeling ... Sample III

Piao et al., Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter, HT’17.
Implicit User Interest Modeling ... Sample III

Piao et al., Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter, HT’17.
Implicit User Interest Modeling ... Sample III

Piao et al., Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter, HT’17.
Implicit User Interest Modeling ... Sample III

Piao et al., Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter, HT’17.
Category-based propagation

Property-based propagation

Piao et al., Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter, HT'17.
Implicit User Interest Modeling ... Sample IV

- Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.
I. Tweets are annotated by TAGME
II. LDA is applied to extract topics and user topic profiles
   i. All tweets of a user form a single document
Implicit User Interest Modeling ... Sample IV

- Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.
Implicit User Interest Modeling ... Sample IV

Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.

Network Representation

Topic graph

User-topic graph

User graph
User-topic graph is a weighted *undirected* graph based on explicit user interests.

Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.
Topic graph is a *weighted undirected* graph based on the inter-topic relatedness measures:

1. Semantics relatedness
2. Collaborative relatedness
3. Hybrid relatedness
User graph is a *weighted undirected* graph based on the co-retweet similarity between the users.

Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.
Implicit User Interest Modeling ... Sample IV

Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018.

Social posts

1. Tweet Entity Linking
2. Topic Interest Modeling

Explicit Interest Detection

3. Network Representation
4. Heterogeneous Link Prediction

Implicit Interest Inference

5. Interest Profile Modeling

Interest profile

PathPredict (Sun et al., PVLDB 2011)
Supervised
meta-paths based
Implicit User Interest Modeling ... Sample IV

Zarrinkalam et al., Mining user interests over active topics on social networks, IP&M 2018

1. Tweet Entity Linking
2. Topic Interest Modeling
3. Network Representation
4. Heterogeneous Link Prediction
5. Interest Profile Modeling

Social posts

Explicit Interest Detection

Implicit Interest Inference

Interest profile
• Introduction

• User Interest Modeling
  • Information Sources
    • Internal
    • External
  • User Interest Representation Units
    • keyword-based
    • Concept-based
  • User Interest Modeling Types
    • Explicit
    • Implicit
User Interest Modeling Type

- Explicit
- Implicit
- Future
User Interest Modeling Type

- Explicit
- Implicit
- Future
Future User Interest Modeling

Allows future planning
- E-commerce business benefits
- Targeted advertising
- Efficient delivery of services

E.g. It is already shown that the box-office revenues of movies can be successfully forecasted in advance of their release by analyzing users' interest in social networks (Asur and Huberman, 2010)

If the forecasted box-office revenues are below expectations, decision makers can provide film promotion in time for coming up to their expectations.
Future User Interest Modeling: Temporality

Users’ interests change over time

(Ahmed et al., KDD’11)
Future User Interest Modeling: Temporality

• Constraint-based approaches
  • Extract user interests within a short-term period or temporal patterns
    • E.g. weekends, last two weeks, 100 recent post, last 90 days (Spasojevic et al. KDD’14)

• Interest decay functions
  • The weights of user interests were discounted by time, i.e., the interests appearing a long time ago would decay heavily
    • E.g. exponential decay function (Bao et al., DSSs 2013)
    • Time-sensitive interest decay function (Abel et al. 2011)
Predict the degree of users’ interest in future over a set of observed topics

- Users’ interests change over time
- Users’ behaviors are affected by opinions of their friends
- Set of topics: Trending topics in Sina-weibo

(a) Social network  (b) $R_t$: user-topic matrix in $T_1$  (c) $R_t$: user-topic matrix in $T_2$  (d) $R_t$: user-topic matrix in $T_3$  (e) $R_t$: user-topic matrix in $T_4$
Bao et al., A new temporal and social PMF-based method to predict users' interests in micro-blogging, DSSs 2013
Future User Interest Modeling ... Sample I

- Bao et al., A new temporal and social PMF-based method to predict users' interests in micro-blogging, DSSs 2013

SocialMF \( \text{(Jamali and Ester, RecSys '10)} \)

User-Topic Matrix

Users’ Social Network
Future User Interest Modeling … Sample I

Bao et al., A new temporal and social PMF-based method to predict users' interests in micro-blogging, DSSs 2013
Predict the degree of users’ interest in future over a set of observed topics

- Users’ interests change over time
- Users' behaviors are affected by opinions of others!

Based on *Granger causality* and for topic $z$, a user $c$ (cause) influences another user $e$ (effect), i.e., $c \rightarrow^G_z e$ if the past observations of $c$ lead to a more accurate prediction of the behavior of $e$ toward $z$ compared to when the past observation of $e$ is considered alone.

User-topic timeseries of two users $c$ and $e$ for topic $z$
Future User Interest Modeling ... Sample II

Arabzadeh et al., Causal Dependencies for Future Interest Prediction on Twitter, CIKM’18.

1. Temporal user interest modeling
2. Influencer Identification
3. User Interest Prediction

Social Posts ➔ Temporal user interest modeling ➔ User-topic Timeseries ➔ Influencer Identification ➔ Influence Network ➔ User Interest Prediction ➔ Future User Profile
Future User Interest Modeling ...

- Arabzadeh et al., Causal Dependencies for Future Interest Prediction on Twitter, CIKM’18.

**Social Posts**

1. Temporal user interest modeling
2. Influencer Identification
3. User Interest Prediction

- LDA on daily tweets of a user as single document
- Daily user-topic timeseries for each topic of interest $z$: $X_{ez} = [x_{ez,1}, x_{ez,2}, \ldots, x_{ez,T}]$
Future User Interest Modeling ... Sample II

Arabzadeh et al., Causal Dependencies for Future Interest Prediction on Twitter, CIKM’18.

For each topic $z$, an influence network is represented by a weighted directed user graph whose edges’ weights are based on Granger causality.
Given a user and a topic, her future interest is calculated by Vector AutoRegression (VAR) using her Top-k influencers.
Future User Interest Modeling

In prior work, it is assumed that the set of topics stay the same over time

- Unrealistic assumption in social networks where topics can rapidly change in reaction to real world events
- Cannot predict user's interests with regard to new topics since these topics have never received any feedbacks from users in the past.
- Does not address cold item problem
Future User Interest Modeling … Sample III


User interest prediction over future *unobserved* topics

- Users’ interests change over time
  - Topics of interest rapidly change in reaction to real world events

- Intuition: although the users’ topics of interest change over time, users tend to incline towards topics and trends that are semantically or conceptually similar to a set of core interests.
  - They model high-level interests of users by utilizing semantic information from knowledge bases such as Wikipedia.
Future User Interest Modeling ... Sample III

I. Tweets are annotated by TAGME
II. Time period is broken into time intervals $t$
III. In each item interval $t$:
   i. All tweets of a user form a single document
   ii. LDA is applied to extract topics and user topic profiles

Future User Interest Modeling ... Sample III


Hierarchical Category Profile Identification
## Summary: Information Source

<table>
<thead>
<tr>
<th>Work</th>
<th>Internal</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Social Posts</td>
<td>User Relations</td>
</tr>
<tr>
<td><strong>Explicit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Liang et al., KDD’18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Abel et al., ESWC’11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Implicit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., IP&amp;M 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Trikha et al. ECIR’18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Faralli et al., J. Web Semantics 2017)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Piao and Breslin, ECIR’17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Piao and Breslin, HT’17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., ECIR’16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(He et al., CIKM15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Wang et al., TIST’2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Kapanipathi et al., ESWC’14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Future</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., IRJ 2019)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Arabzadeh et al., CIKM’18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Bao et al., DSSs 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work</td>
<td>Keyword</td>
<td>Group of keywords</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Explicit</strong></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Liang et al., KDD’18)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Kang and Lee, J. Info. Sys. 2017)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Abel et al., ESWC’11)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td><strong>Implicit</strong></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., IP&amp;M 2018)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Trikha et al. ECIR’18)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Faralli et al., J. Web Semantics 2017)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Piao and Breslin, ECIR’17)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Piao and Breslin, HT’17)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., ECIR’16)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(He et al., CIKM’15)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Wang et al., TIST 2014)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Kapanipathi et al., ESWC’14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Future</strong></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., IRJ 2019)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Arabzadeh et al., CIKM’18)</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>(Bao et al., DSSs 2013)</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>
## Summary: Approaches

<table>
<thead>
<tr>
<th>Work</th>
<th>Frequency based</th>
<th>Probabilistic Prediction</th>
<th>Link Embedding</th>
<th>Regression</th>
<th>Frequent Pattern Mining</th>
<th>Matrix Fact.</th>
<th>Spreading Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit</td>
<td>(Liang et al., KDD’18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Abel et al., ESWC’11)</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implicit</td>
<td>(Zarrinkalam et al., IP&amp;M 2018)</td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Trikha et al. ECIR’18)</td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| | (Faralli et al., J. Web Semantics 2017) | | * | | | * | *
| | (Piao and Breslin, ECIR’17) | | * | | * | | *
| | (Piao and Breslin, HT'17) | | * | | | * | *
| | (Zarrinkalam et al., ECIR’16) | | * | | * | | *
| | (He et al., CIKM’15) | | * | | | | *
| | (Wang et al., TIST 2014) | | * | | | | *
| | (Kapanipathi et al., ESWC’14) | | * | | | * | *
| Future | (Zarrinkalam et al., IRJ 2019) | | * | * | | | *
| | (Arabzadeh et al., CIKM’18) | | * | * | | | *
| | (Bao et al., DSSs 2013) | | * | | | | * |
2 User Interest Modeling

- Information Sources
- User Interest Representation Units
- User Interest Modeling Types
- User Interest Datasets
- Evaluation Methodologies
User Interest Dataset

• A comprehensive dataset that contains user posts, social relations and reliably extracted user interests is required for evaluating user interest modeling approaches.

• Due to the challenges of data collection from social networks and reliably detecting users interests there is no comprehensive user interest dataset.
  • A recent work provides a large multi-domain user interests dataset of Twitter which is useful in this context (Tommaso et al., ISWC’18)
Map interest to wikipages

Map topical followee to wikipages, e.g., WIKI:EN:Arsenal_F.C.

Identify topical followee as interest, e.g., @Arsenal
+ Non-reciprocal
+ High in-degree
+ Verified Accounts

Extract interest from tweets
Stream tweets including #tag and related URL

Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.
Try to *reliably* extract user interests from tweets using
- Pre-formatted messages (e.g., for Spotify: NowPlaying <title> by <artist> <URL>)
- Service-mediated generation of tweets including the URL (e.g., for Spotify #NowPlaying followed by the URL of music)
  - In English: *Spotify* for music, *IMDb* for movie, and *goodreads* for book
  - In Italian: *Spotify* for music, *TvShowTime* for movie, and *aNobii* for book
Mapping interests extracted from users’ messages to Wikipedia pages is a very reliable process, given the additional contextual information extracted from the URL.

- **Music**: <Title, Author (eg. singer, band)>
- **Books**: <Title, Author>
- **Movie**: <Title, Year of production, Type (eg. movie, tv series)>

Extract interest from tweets

Stream tweets including #tag and related URL

Map interest to wikipages

---

Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.
Users' interests can also be extracted from the authoritative (topical) friends

- Less volatile and relatively stable indicators of a variety of interests
  - Users tend to be stable in their relationships (Myers and Leskovec, WWW’14)
  - Different domains, such as entertainment, sport, art and culture, politics, etc.

Topical friends

- Mostly are not reciprocated
- Have a high in-degree
To distinguish between topical and peer friends, boolean SVM classifier is trained:

- In degree
- Out degree
- In/out degree ratio
- Binary feature to indicate presence of words such as singer, writer, ... in the user’s Account profile

Positive samples are Verified Twitter Accounts as authentic accounts of public interest.
User Interest Dataset ... Sample I

- Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.

Identify *topical* followee as interest, e.g., @Arsenal

- Non-reciprocal
- High in-degree
- Verified Accounts

Map topical followee to wikipages, e.g., WIKI:EN:Arsenal_F.C.
Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.

Mapping topical friends to Wikipedia is rather complex

- Twitter profile is misleading
- Twitter profile does not provide sufficient context

Bill Gate's in his Twitter profile: “Sharing things I’m learning through my foundation work and other interests...

Bill Gate’s in Wikipedia article: “William Henry Gates III (born October 28, 1955) is an American business magnate, ...“

**Ensemble Voting** on max agreement (M1,M2,M3,ESM1,ESM2,ESM3) with at least 2 in agreement.
### User Interest Dataset ... Sample I

Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.

#### Statistics

<table>
<thead>
<tr>
<th></th>
<th>6-months (April-September 2017)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of unique interests</td>
<td>Avg interests/user</td>
</tr>
<tr>
<td><strong>English speaking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>users</td>
<td>[U] = 444,744</td>
<td></td>
</tr>
<tr>
<td>message-based</td>
<td>282,303</td>
<td>6</td>
</tr>
<tr>
<td>topical friend-based</td>
<td>58,789</td>
<td>82</td>
</tr>
<tr>
<td><strong>Italian speaking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>users</td>
<td>[U] = 25,135</td>
<td></td>
</tr>
<tr>
<td>message-based</td>
<td>14,895</td>
<td>6</td>
</tr>
<tr>
<td>topical friend-based</td>
<td>4,580</td>
<td>41.96</td>
</tr>
</tbody>
</table>

The users’ interests are stored as RDF using SIOC & SKOS ontologies.
**User Interest Dataset ... Sample I**

- Tommaso et al., Wiki-MID: a very large Multi-domain Interests Dataset of Twitter users with mapping to Wikipedia, ISWC’18.

<table>
<thead>
<tr>
<th>Source</th>
<th>Interest</th>
<th>Wikipage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>Eyes Wide Open - 2009 - movie</td>
<td>WIKI:EN:Eyes_Wide_Open_(2009_film)</td>
</tr>
<tr>
<td></td>
<td>Okja - 2017 - movie</td>
<td>WIKI:EN:Okja</td>
</tr>
<tr>
<td>Goodreads</td>
<td>The Beautifull Cassandra - Jane Austen</td>
<td>WIKI:EN:Jane_Austen</td>
</tr>
<tr>
<td></td>
<td>The Beach - Alex Garland</td>
<td>WIKI:EN:The_Beach_(novel)</td>
</tr>
<tr>
<td>Spotify</td>
<td>I Don’t Know What I Can Save You From - Kings of Convenience</td>
<td>WIKI:EN:Kings_of_Convenience!</td>
</tr>
<tr>
<td></td>
<td>Nothing Matters When We’re Dancing - The Magnetic Fields</td>
<td>WIKI:EN:The_Magnetic_Fields</td>
</tr>
<tr>
<td>Topical friends</td>
<td>@IMDb</td>
<td>WIKI:EN:IMDb</td>
</tr>
<tr>
<td></td>
<td>@UNICEF_uk</td>
<td>WIKI:EN:UNICEF_uk</td>
</tr>
<tr>
<td></td>
<td>@TheMagFields</td>
<td>WIKI:EN:The_Magnetic_Fields</td>
</tr>
<tr>
<td></td>
<td>@BarackObama</td>
<td>WIKI:EN:Barack_Obama</td>
</tr>
<tr>
<td></td>
<td>@Spotify</td>
<td>WIKI:EN:Spotify</td>
</tr>
</tbody>
</table>
User Interest Modeling

- Information Sources
- User Interest Representation Units
- User Interest Modeling Types
- User Interest Datasets
- Evaluation Methodologies
Evaluation Methodologies

- Intrinsic
- Extrinsic
Intrinsic: *Is it good in and of itself?*

- User Study
- Labeled Dataset
- Qualitative Analysis

Extrinsic
Collecting explicit feedback from users about inferred interests

I. Target User

- The most direct and accurate evaluation methodology
- Small-scale: small number of users are willing to participate in the user study
  - E.g., 30 out of 500 users (Budac et al., AAAI’14)

II. Third Party

- Large-scale (possibly for millions of users for a nominal cost)
  - Using crowdsourcing platforms such as Amazon Mechanical Turk platform
- Identifying latent interests of some other user based on her published posts is a hard task for human beings and the quality of evaluation results would be questionable at best!
Inferred interests of users are compared with their real interests

► Help with supervised user interest modeling

I. Annotated by human annotators
   ► Small-scale: e.g., 20 Facebook users and 70 Twitter (Kang et al., JIIS 2019)

II. Using explicit interests of users as ground truth
   E.g., extract the explicit interests of users from their biography based on some predefined patterns using the Stanford POS-Tagger (He et al., CIKM ’15)
   “play X”, “X fan”, “interested in X”, “love X”

► The quality of evaluation results would be questionable!
Intrinsic Evaluation Methodology: Qualitative

Present some representative user interest profiles (e.g., politicians, researchers and technology news), and discuss the quality of profiles.

• Usually used together with other evaluation methods

<table>
<thead>
<tr>
<th>User, with extracts from bio</th>
<th>Top topics of interest, inferred by different methodologies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>List-topics (proposed)</td>
</tr>
<tr>
<td>Michelle Zhou: into interior design, love shopping &amp; food ...</td>
<td>interior design, decor, fashion, shopping, lifestyle, travel, drinks, hotel</td>
</tr>
<tr>
<td>Erin Marshall: sharing my life in fashion, fun ... Love red lipstick, high heels</td>
<td>fashion, fun, style, fashion designers, beauty</td>
</tr>
<tr>
<td>Jesse Millar: Computer sc. student, addicted to programming, game design</td>
<td>developers, game dev, technology, games</td>
</tr>
<tr>
<td>Mattia Pontacolone: loves social media, mobile, web apps and start-ups</td>
<td>technology, social media, marketing, startup,</td>
</tr>
<tr>
<td>Cege Smith: Obsessed with The Vampire Diaries, Cult, and Being Human. Author ...</td>
<td>authors, publishing, writing, vampirediaries, writing resources</td>
</tr>
</tbody>
</table>

(Bhattacharya et al. RecSys’14)
Intrinsic

Extrinsic: Does it help to solve or improve another problem?
Extrinsic Evaluation Methodology

Evaluation in terms of the performance of a specific application

- Without imposing an extra burden on users
- Facilitates the comparison of different user modeling strategies

- Does not directly evaluate the inferred user interest profiles, by the opinions of users
- The evaluation is in the context of a specific application
  - Different user interest profiles have different levels of performance on different applications.
Extrinsic Evaluation Methodology

Some sample applications for extrinsic evaluation:

- News recommendation
- URL recommendation
- Friend recommendation
- Retweet prediction
- ...

For each application we need to answer:

- How to obtain the known output of the application? (i.e., Ground truth)
- How to incorporate user interests profiles in the application? (i.e., Recommendation algorithm)

By comparing the results of the application with the ones in the ground truth, it is possible to evaluate the quality of the results, and therefore determine how successfully the interests of a user have been identified.
## Extrinsic Evaluation Methodology

<table>
<thead>
<tr>
<th>News recommendation</th>
<th>Friend Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Abel et al., UMAP ’11)</td>
<td>(Han and Lee, J. Information Science 2016)</td>
</tr>
</tbody>
</table>

### Ground truth
For each user, the news articles from BBC or CNN to which the user has explicitly linked in her tweets (or retweets) by mentioning the corresponding URL

### Recommendation algorithm
1. Each news article is represented using the same representation model (e.g. a weighted vector over Wikipedia concepts)
2. Ranks the candidate news based on their cosine similarity to the user interest profile

### Ground truth
For each user, her followees are considered as her friends

### Recommendation algorithm
Ranks the candidate user based on computing the cosine similarity between their interest profile and the interest profile of the target user
Evaluation Metrics

• The usual IR metrics:
  • Precision
  • Recall
  • F-Measure

• Metrics for evaluating ranking quality:
  • Mean Reciprocal Rank (MRR): at which rank the first item relevant to the user occurs on average
  • Mean Average Precision (MAP): how well the interests are ranked at top-k and how early relevant results appear
  • Success at rank N (S@N): the mean probability that relevant item occurs within the top-N ranked
  • Recall at rank N (R@N): the mean probability that relevant items are successfully retrieved within the top-N recommendations
  • Precision at rank N (P@N): the mean probability that retrieved items within the top-N recommendations are relevant to the user
Evaluation Metrics

• Metrics for evaluating the prediction of user interests:
  • Mean Absolute Error (MAE)
  • Root Mean Square Error (RMSE)

• Metric to evaluate the overall generalization ability of modeling unseen or held-out data
  • Perplexity
    • A low perplexity indicates better performance.
    • For each user, the interests of users as a topic distribution is modeled based on a given time interval and her tweets in the next time interval is considered as the held-out document
<table>
<thead>
<tr>
<th>Work</th>
<th>Methodology</th>
<th>Dataset</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explicit</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Liang et al., KDD’18)</td>
<td>*</td>
<td>Twitter,</td>
<td>P@N, nDCG, MRR, MAP</td>
</tr>
<tr>
<td>(Abel et al., ESWC’11)</td>
<td>*</td>
<td>Twitter,</td>
<td></td>
</tr>
<tr>
<td>(Trikha et al., ECIR’18)</td>
<td>*</td>
<td>Twitter,</td>
<td>MAP</td>
</tr>
<tr>
<td>(Zarrinkalam et al., IP&amp;M 2018)</td>
<td>*</td>
<td>Twitter,</td>
<td></td>
</tr>
<tr>
<td>(Faralli et al., J. Web Semantics 2017)</td>
<td>*</td>
<td>Twitter,</td>
<td>Perplexity, MAP, nDCG</td>
</tr>
<tr>
<td>(Piao and Breslin, ECIR’17)</td>
<td>*</td>
<td>Twitter,</td>
<td>MRR, S@N, P@N, R@N</td>
</tr>
<tr>
<td>(Piao and Breslin, HT’17)</td>
<td>*</td>
<td>Twitter,</td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., ECIR’16)</td>
<td>*</td>
<td>Twitter,</td>
<td>AUROC, AUPR</td>
</tr>
<tr>
<td>(He et al., CIKM15)</td>
<td>*</td>
<td>Twitter,</td>
<td>Recall</td>
</tr>
<tr>
<td>(Wang et al., TIST 2014)</td>
<td>*</td>
<td>Twitter,</td>
<td>MRR, P@N, Perplexity</td>
</tr>
<tr>
<td>(Kapanipathi et al., ESWC’14)</td>
<td>*</td>
<td>Twitter,</td>
<td>MRR, Precision</td>
</tr>
<tr>
<td><strong>Implicit</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Zarrinkalam et al., IRJ 2019)</td>
<td>*</td>
<td>Twitter,</td>
<td>MAP, nDCG</td>
</tr>
<tr>
<td>(Arabzadeh et al., CIKM’18)</td>
<td>*</td>
<td>Twitter,</td>
<td>MAP, nDCG, P@N</td>
</tr>
<tr>
<td>(Bao et al., DSSs 2013)</td>
<td>*</td>
<td>Sina-Wbio,</td>
<td>P@N</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. User Interest Modeling
3. Challenges and Future Directions
More Semantics I

The fundamental step in concept-based user interest modeling is challenging by itself, i.e., Extracting entities from

- social posts | biographies | list memberships | usernames

Accuracy of the annotator has an influence on the accuracy of the user interest model

- Most work overlook the uncertainty (confidence score) in annotators

More research on the impact of using annotators in the user interest modeling is needed!
Existing work has mainly rely on explicit entity annotators

“The movie Gravity was more expensive than the Mars Orbiter Mission.

► There can be many entities implicitly mentioned in posts!

“This new space movie [Gravity] is crazy. you must watch it!

21% of the entities mentioned in the Movie domain and 40% in the Book domain are implicit entities (Perera et al., ESWC ’16).

More research on the potentiality of implicit entities in the user interest modeling domain is needed
Majority of the current work in user interest modeling is descriptive in nature, i.e., they find the *current* interests of a user.

- What topics a user will be interested in or attracted to in the future?
- How would a user react if a given brand released a new product in near future?

*Lack of predictive and prescriptive insight*
More Explainability

The explainability of recommender systems has attracted considerable attention for two main goals:

- Transparency: helping users to understand how the system works
- Scrutability: allowing users to tell the system if it is wrong

Taking explainability to the level of user interests (Balog et al., SIGIR ‘19)

- You like movies that are tagged as ‘action’, especially those that are tagged as ‘violent’, such as Aliens.
- You like movies that are tagged as ‘twist ending’, such as A Pure Formality.
- You don’t like movies that are tagged as ‘adventure’, unless they are tagged as ‘thriller’, such as Twister.
- You like movies that are tagged as ‘cheesy’, such as Who Framed Roger Rabbit?
- You like movies that are tagged as ‘australia’, such as Crocodile Dundee II.

Lack of explainability in user interest modeling from OSNs
Existing work on user interest modeling has mainly focused on a single social media due to the challenges of data integration between different social media.

► It is crucial to integrate information from *multiple* social media to provide a 360° view of the user’s behavior.
  - On average, people actively use ~3 social media platforms and this value is increasing (GlobalWebIndex).

New technologies have been recently developed to link different social media accounts of the same user (Shu et al., KDD’17).
More Dynamicity I

It has been well-accepted that users’ degree of interest toward topics are dynamic. However, social networks have more dynamicity in other aspects as well such as:

• Users join | leave
• Topics emerge | disappear
• Inter-user interaction changes
• Inter-topic relatedness changes

A unified model is a requirement to incorporate all dimensions of dynamicity which is yet to be explored.
Although it is already shown that modeling social network information as a **unified graph** and applying **heterogeneous** link prediction methods is promising to predict user interests (Zarrinkalam et al., IP&M 2019).

- The dynamicity is overlooked.
An approach that consider both heterogeneity and dynamicity in a unified model is needed.

Relationship Prediction in Dynamic Heterogeneous Information Networks (Fard et al., ECIR ‘19)
Recently, dynamic network embedding has shown promising result in underlying tasks such as:

- **Node Classification** (Cavallari et al., CIKM’17)
- **Link Prediction** (Grover and Leskovec, KDD’16)
- **Community Detection** (Perozzi et al., KDD’14)

Dynamic network embedding could be a promising approach for predicting users’ interests, but they fall short in heterogeneous networks.

*Graph embedding methods which are able to incorporate dynamicity and heterogeneity is needed.*
More Embedding II

Neural network based embedding methods have shown their power in learning social media-based user embedding \cite{PangDing2019}:

- Word2Vec | Doc2Vec | RNN | LSTM

User embedding features are latent features automatically uncovered by the system, i.e., not interpretable.

- Difficult for humans to understand the meaning of these features
- It may significantly impact the ability to gain insight into user behavior

More Study on applying neural network-based embedding to produce interpretable user interests models is needed \cite{Liang2018}.
Dynamic profile of users using *relevant* & *diversified* keywords

Liang et al., Dynamic Embeddings for User Profiling in Twitter, KDD’18.

$U_t, D_{\leq t} \rightarrow \text{Dynamic embedding of users & words in *same* space (DUWE)} \rightarrow \text{Word diversification (SKDM)} \rightarrow W_t$
“Which combinations of different approaches in each dimension can provide the best user interest profiles in each application?

- **URL recommendation** (Piao and Breslin, EKAW’16)
  - The most important factors:
    - WordNet synsets and DBpedia entities, and content enrichment by the content of URLs
    - Interest decay functions perform better than constraint-based approaches such as short- and long-term profiles.

- **Publication Recommendation** (Nishioka and Scherp, iKNOW’16)
  - Constraint-based approach outperforms exponential decay function

*More study on the effect of different user modeling dimensions on different applications.*
Different datasets are used in different studies which makes the comparison difficult. No common benchmarks and datasets is available.

Recently, a user interests dataset is published (Tommaso et al., ISWC 2018)

- Multi-domain preferences (music, books, movies, celebrities, sport, …)
- 0.5 million Twitter users, April-September 2017
- Dynamicity of users’ interests is overlooked
- No social structure

More study on preparing benchmarks and datasets for user interest modeling is needed
More Reproducibility

- Implementation for most of the proposed methods are not publicly available.

- No library exists including different user interest modeling approaches
  - E.g. LibRec (Java library for recommender systems)
• Introduction

• User Interest Modeling
  • Information Sources
  • User Interest Representation Units
  • User Interest Modeling Types
  • User Interest Dataset
  • Evaluation strategies

• Challenges and Future Directions
Organizers

Fattane Zarrinkalam
Postdoctoral Fellow Laboratory of Systems, Software and Semantics (LS3), Ryerson University, Canada
@FattaneZ
fzarrinkalam@ryerson.ca
http://ls3.rnet.ryerson.ca/people/fattane_zarrinkalam/

Hossein Fani
Ph.D. Candidate at University of New Brunswick
@HosseinFani
hfani@unb.ca
https://sites.google.com/site/hosseinfani/

Ebrahim Bagheri
Associate Professor & Director for the Laboratory for Systems, Software and Semantics (LS3)
Ryerson University, Canada
Canada Research Chair in Software and Semantic Computing
NSERC/WL Industrial Research Chair in Social Media Analytics
@ebrahim_bagheri
bagheri@ryerson.ca
http://ls3.rnet.ryerson.ca/