Understanding User Behavior From Online Traces

ELAD KRAVI

ADVISORS

YARON KANZA

BENNY KIMELFELD

The Data Revolution

People share large amount of data

- Explicitly and implicitly
- Attributes collected including
  - locations, timestamps, textual content etc.

A great opportunity to *improve* online services, to *enhance* existing infrastucture and to *engage* users

Goals

Leverage analysis of online traces for

- Improving measurement of *users' similarity*
- Enhancing *online services*
- Engaging *online activity*

What affects users online behavior?

- Do people have *different needs* in *different places*?
- How do *social relationships* affect online behavior?

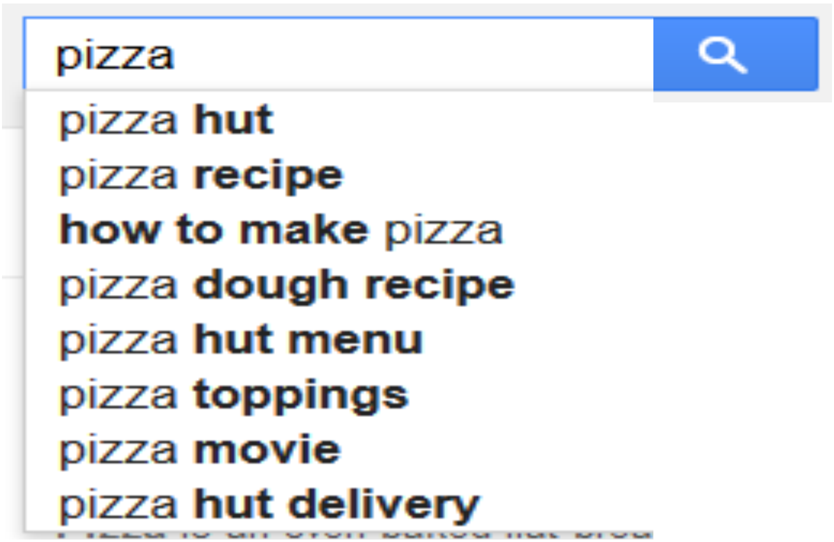
Outline

- Location and text effect
- Social networking effect

Location and Text Effect

Location	The *City Nexus* tool [SIGSPATIAL 2014]
Textual	Multi-Clicked Queries [under review]
Location + Textual	Familiarity of environment [SIGIR 2015] Next
	Correlation Between Textual Content and Geospatial Locations [GeoRich 2014]

Does the familiarity of environment matter?



- **Pizza dough?**
- **Pizza place?**

A note on the dataset

Our dataset included more than a billion queries log traces of a popular commercial search engine

Using these traces one can calculate the rank of query auto-completion completion terms

Query	#
”pizza dough”	5
”pizza place”	3

Query	auto-completion	rank
pizza	dough	1
	place	2

Hypothesis: Information need is affected by *familiarity* of the environment

Category	Familiar	Unfamiliar
pizza	dough:3, places:5	places:3, dough:5
gas	fireplace:3, station:6	station:3, fireplace:8
wild	rice:2, horse:5	horse:2, rice:4

What is a ***location***?

- Using the IP address

What is a ***familiar*** location?

- Significance
- Travels

How to ***verify*** that the model works?

A ***place*** p is ***familiar*** to a ***user*** u if:

% of days on which *u* was active from *p* ≥ t

p is significant to *u*

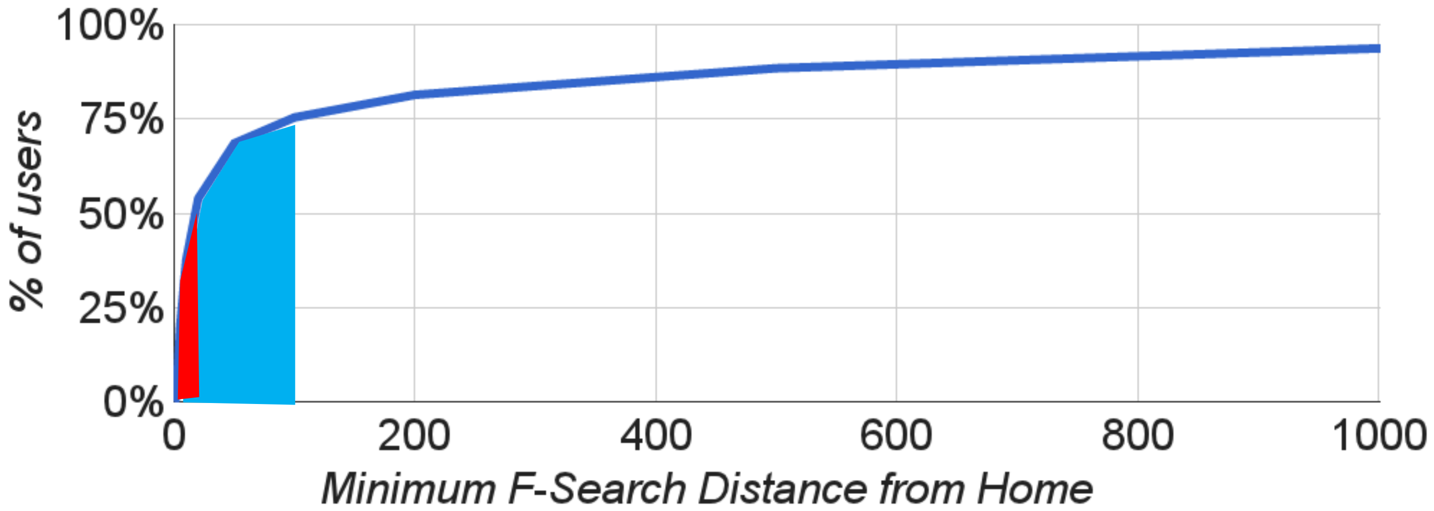
AND

(u returned to p at least r times **OR** All u's activities are in p)

re-occurrence

e.g. desktop

Distance from Declared Home



For **53.9% of the users**, the distance from declared home was **smaller than 20** KMs

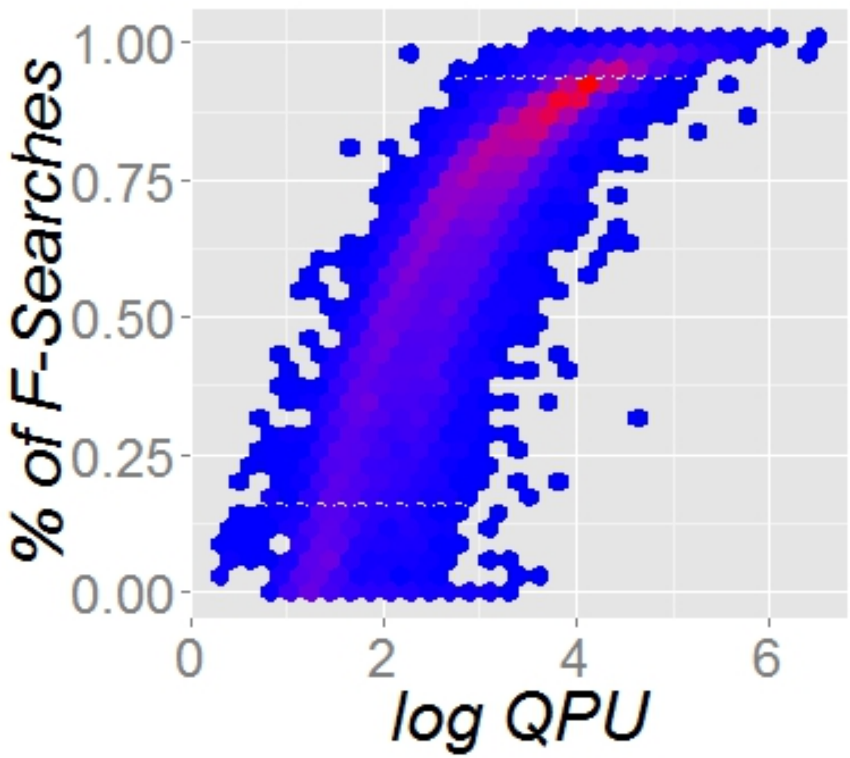
For **75.4% of the users** it was **smaller than 100 kilometers**

Queries Per Unique Users (QPU)

A place having ***large QPU*** (many queries few users) is expected to obtain ***many familiar queries***

- and vice versa

Queries Per Unique Users (QPU)

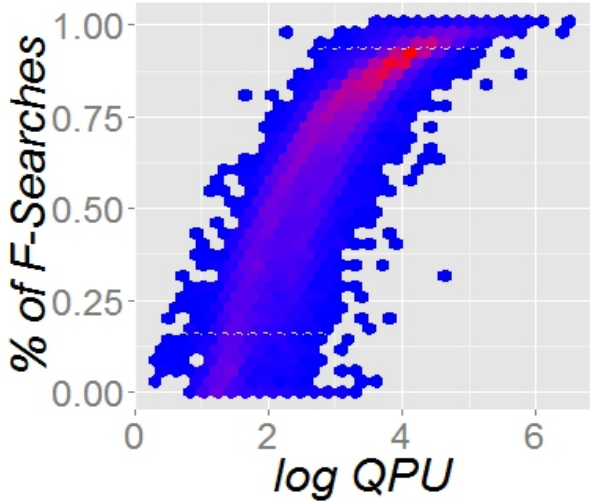


A clear correlation is observed

Compare To A Baseline Model

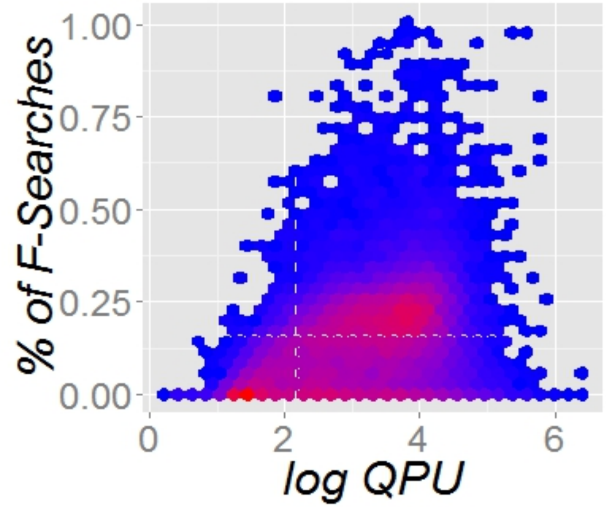
Consider the following baseline:

- A ***familiar*** location is <u>every place around 20 KM from declared home</u>



Our Model

A clear correlation can be seen



20 KM Model

No correlation can be seen

Difference in Language Models

Uni-grams	
***F*-Search**	***U*-Search**
facebook	google
sale	restaurant
free	schedule
games	football
ebay	ny
how	lyrics
login	ct
online	store
craiglist	movie
recipes	hours
porn	locations
tube	mall

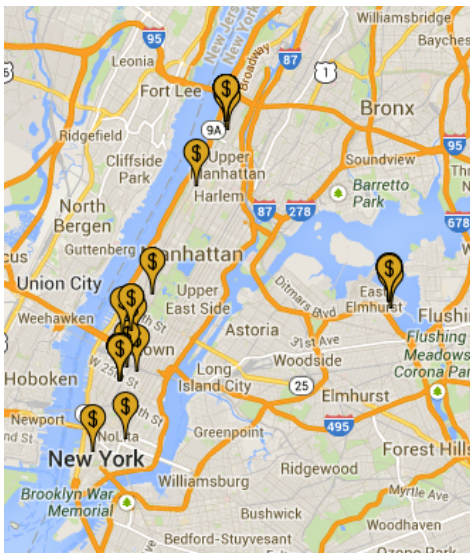
Bi-grams	
***F*-Search**	***U*-Search**
for sale	new york
how to	phone number
facebook login	google search
to make	new jersey
homes for	high school
cool math	how many
you tube	hobby lobby
sales in	in new
funeral home	football schedule
real estate	r us
black friday	movie theater
for kids	nfl scores

Location and Text Effect

Location	The *City Nexus* tool [SIGSPATIAL 2014]
Textual	Multi-Clicked Queries [under review]
Location + Textual	Familiarity of environment [SIGIR 2015]
	Correlation Between Textual Content and Geospatial Locations [GeoRich 2014]

Next

Posts Origin in Microblogs



geo-tagged

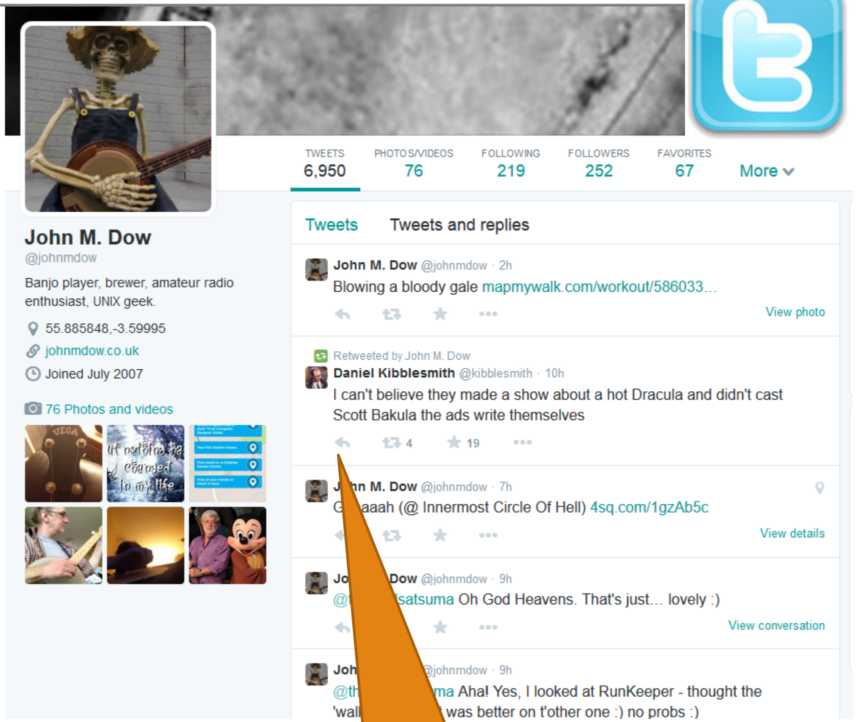
text

Correlation?

Application #1: Associating Posts from Different Networks

Text-based social network

Alice



Location-based social network

Instagram

Bob

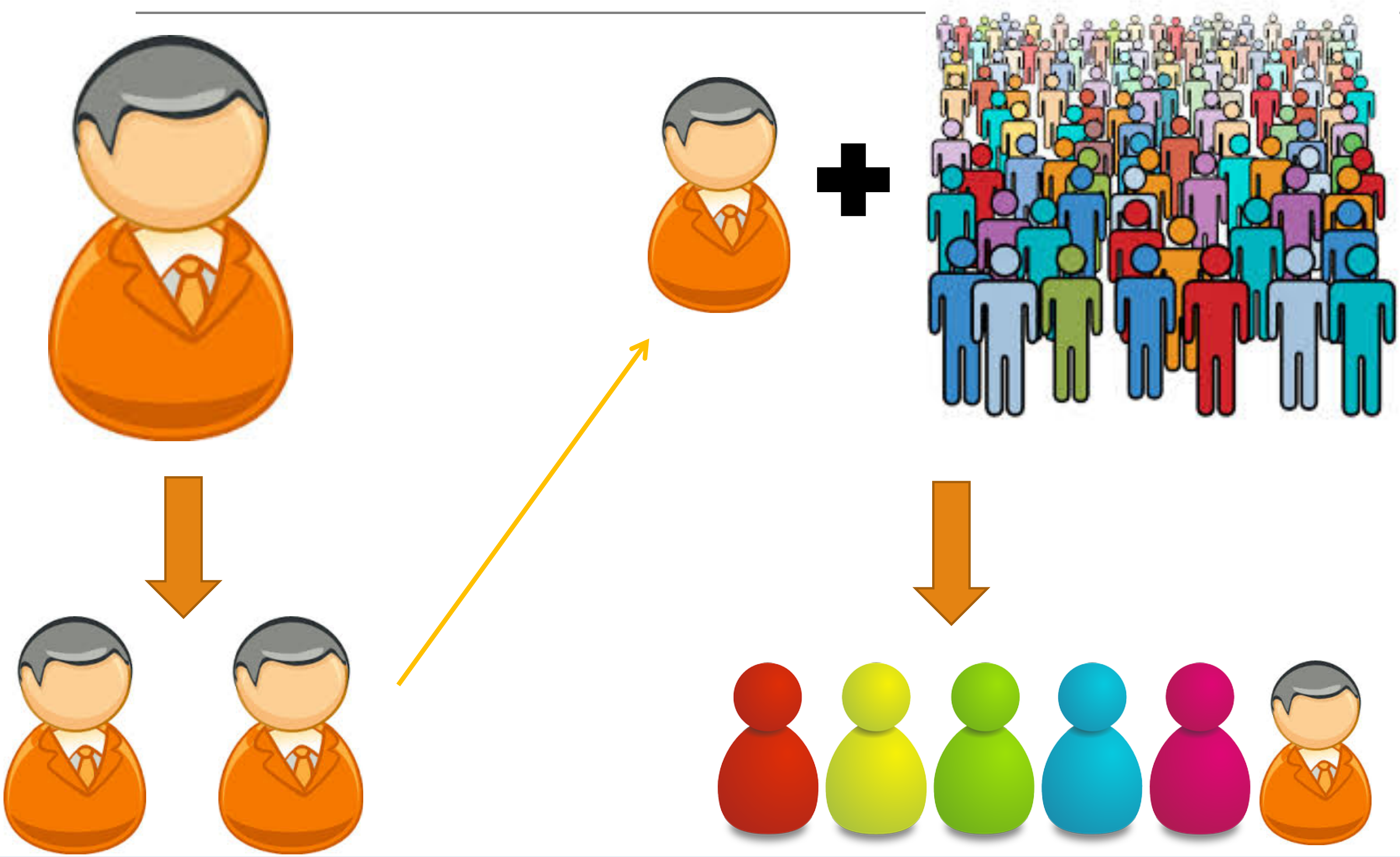
Who will have a greater success, Alice or Bob?

An Example – Measuring User's Similarity

We compared between similarity based on the following measures:

- only the *locations* of the messages using ***nearest neighbor distance***
- only the *content* of the messages using ***TF-IDF***
- combination of **both**

Identification Test



Problem Definition – Identification test

A post p is denoted by $p=(l,c)$

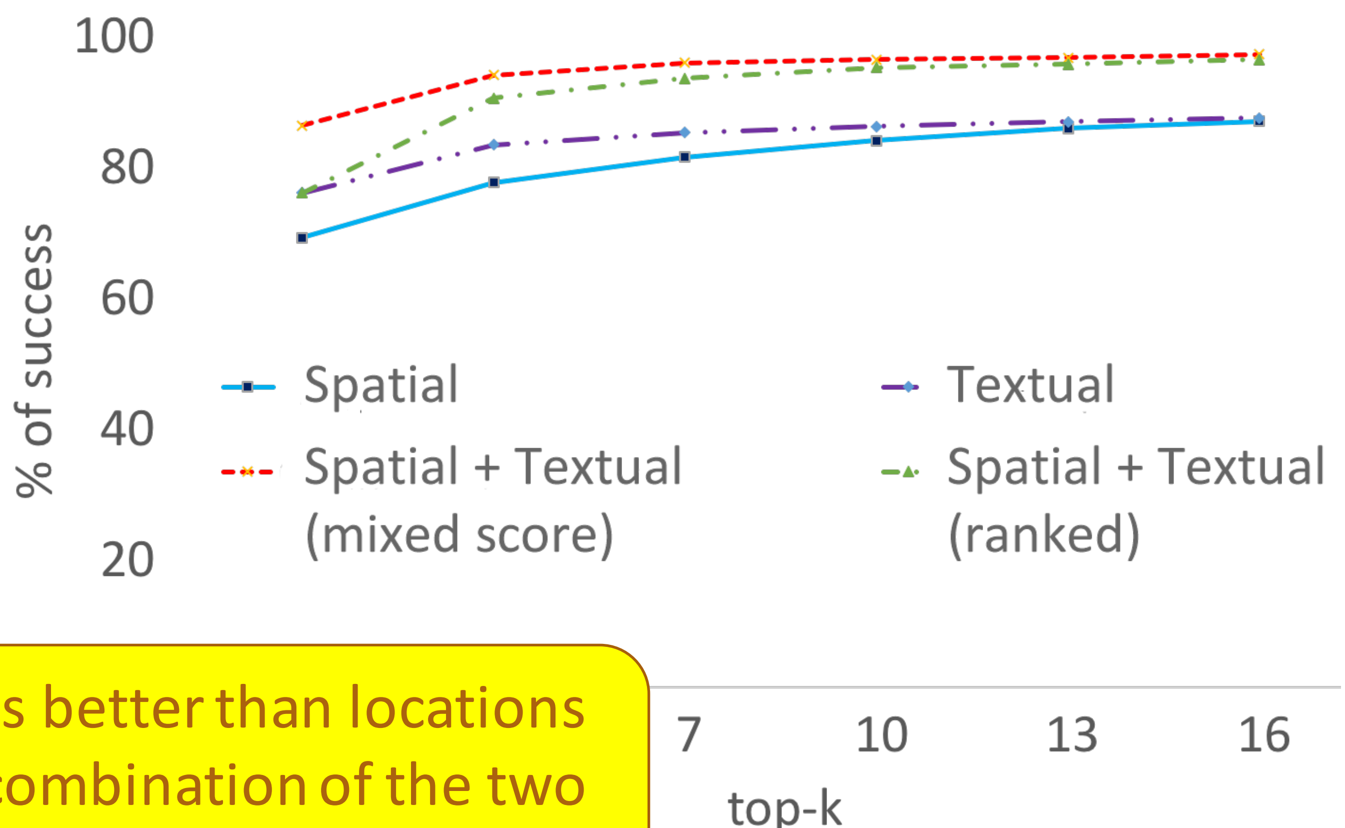
- l – location, on sphere
- c – textual content

Each user u is associated with the set p_u of her posts

- Split u into $\boldsymbol{u_1}, \boldsymbol{u_2}$, such that $p_{u_1} \cup p_{u_2} = p_u$ and $p_{u_1} \cap p_{u_2} = \emptyset$
- Let K be the ***k-most-similar*** users to u_1 among $U \cup u_2$
- Consider success as the case where $u_2 \in K$ and failure otherwise

Goal – maximize success rate

Accuracy as a Function of k



Content is better than locations and the combination of the two provides the best results

Outline

Location and text effect

➡ Social networking effect

Social Networks

Recommending content items to community owners [SIGIR 2014]

- Using recommender-system approach to recommend content items to owners of online communities in a corporate social network

Measuring the effect on activity level [TOCHI 2015]

- Further extending previous work to examine the effect of recommendation over the activity in the communities

Main Challenges

Formal modeling that allows automatic detection

- avoiding detection of erroneous patterns
- yet, portraying the diversity of human behavior

Verifying a proposed **model**

- lack of ground truth and tagged data

Summary

We examined the utilization of spatial, textual and social information toward understanding online behavior

- **Spatial and Textual:** jointly-visited locations, multi-clicks, familiarity of environment and similarity between users
- **Social:** recommending content items to community owners, engaging community's activity

Leveraging online data one can Improve measurement of *users' similarity,* enhance *online services* and engage *online activity*

Future Work

Building tools for finding complex patterns that combine traces from different datasets

- For example - improving web search by using social activity

Developing infrastructure for allowing users to define their models of online behavior patterns, and later on detecting these patterns on real datasets

Thank You!


Technion
Recommending
Trend-Detection
Social-Networks
Search
Big-Data
GEORICH
Unfamiliar
media
Enterprise
Activity
Familiar
Recommendation
SIGIR
Online
SIGMOD
User
Clicks
One
Understanding
Content
Many
owners
Location
Text
using
Strange
TOCHI
Locations
Land
social
Yahoo
SIGSPATIAL
CUP
IBM
Machine-Learning
Temporal
Multiple
Communities
Query
Correlated
Increasing
Web
Users
Geospatial
Searcher
Microblogs
community
Queries
Search-Engine
Natural-Language-Processing
Information-Retrieval
Similarity
content