Keyword Search in Social Networks

Advisor/Committee Members
Dr. Chris Pollett
Dr. Mark Stamp
Dr. Soon Tee Teoh

By
Vijeth Patil
 Agenda

- Motivation
- Project goal
- Background
- Yioop!
- Twitter
- RSS
- Modifications to Yioop!
- Test and Results
- Demo
- Conclusion
Motivation

- More and More data accumulated in Social Networks.
- Lot of important information is shared in the social networks.
- Most content never reaches you or its too much.
- Hard to dig these information when you need it
- It matters!
To enhance the search process by allowing users to simultaneously see web and social search results in Yioop! an Open Source search engine

- Provide results from feeds posted by people followed by or friends of the user in the social network.

- Provide results from Really Simple Syndication (RSS) Feeds subscribed by user.
Background

- Social search - takes into account the content from social graph of the person
- Lot of social content is private and accessible only by the user.
- Needs user authorization to access his social network content.
- Most of the media websites provide RSS web feed formats to publish frequently updated works—such as blog entries and news headlines.
- Major search engines are moving towards social.
Access Control

- Most of the social network content is considered private and visible to the user, friends and followers.

- Need permission from user to access this data.

- Creating application on the social network platform and gathering authorization from user.
Yioop!

- Yioop! is an open source search engine and distributed crawler developed by Dr. Chris Pollett in PHP

- Can be configured for General purpose or Personal web crawl

- Provides web interface for controlling and configuring the crawl.

- Stores the crawl data in web archive file format. Indexes Internet Archive's arc format, Open Directory Project RDF data, Wikipedia xml dumps, etc.
To install Yioop! you need:

- A web server (e.g. Apache)
- PHP 5.3 or higher
- Curl library (for simultaneous download of web pages)
- Yioop installation and configuration documentation can be found at www.seekquarry.com
More on Yioop!

- Yioop! uses a simplified distributed model containing nodes with a **name server** to coordinate between nodes.

- **queue server** is a process that performs indexing and scheduling.

- **fetcher** is responsible for downloading pages.
Web search vs Social search

- **Web search**: search for the incoming request from a set of indexed web pages.

- **Positives**:
  - Very effective in finding the required results.
  - Document centric: Results are ranked based on the importance of page to the query.

- **What’s Missing**:
  - Not fresh data.
  - Does not rank based on users’ known sources.
Web search vs Social search contd...

- Social search: gives more importance to the content created by people in the user’s social graph.
- Takes account of various metadata, such as relationship with user, recency, popularity, etc.
- Web and Social Network data are disjoint. User requires to get best of both in a unified search system.
Twitter

- Free social networking and social blogging service that enables its users to send and read messages known as tweets.
- Tweets are text based posts up to 140 characters displayed on the author's profile page and delivered to the author's subscribers who are known as followers.
- If two users follow each other then they are friends in the network.
Twitter

- Twitter user’s content can be accessed by Twitter applications (consumers) authorized by the user.
- An application can access content by making calls to APIs (providers). Example: REST API
- Twitter uses OAuth for authentication
- Each application has its own unique consumer key and consumer secret.
OAuth is an open standard for authorization.

- Allows secure API authorization in simple and standard method for sharing private data.

- User can grant a third party site access to their information stored with another service provider without sharing their access permissions (credentials) or the full extent of their data.
OAuth authentication flow

Consumer (App)
- Request Request Token
- Direct user to Service provider
- Request AccessToken
- Access protected data

Service Provider (API)
- Grant Request Token
- Obtain user authorization
- Redirect user to consumer
- Grant AccessToken
- Redirect user to consumer
Twitter API

- REST API

- Home Timeline
  - Returns the most recent statuses, including retweets if they exist, posted by the authenticating user and the users they follow.
  - This is the same timeline seen by a user when they login to twitter.com

- Can return maximum of 800 tweets per user in an hour.

- JSON data returned from API needs to be parsed.
Structure of Twitter data

- Request made to Twitter REST API for the home timeline data of the user. API returns the JSON data of the format:

```json
[ {
  "text": "US Military stopped helping 'The Avengers' because the movie was too unrealistic",
  "id": 18700887835,
  "created_at": "Fri Jul 16 16:58:46 +0000 2010",
  ...
  "user": {
    "name": "Daniel Burka",
    "followers_count": 3395,
    "friends_count": 542,
    "following": true,
    "screen_name": "cindyli",
    "verified": false,
    ...
  }
  "retweet_count": 13,
  "source": "web",
  ...
},
{ ...
},
]```

- A sample tweet is shown, including the text, ID, creation date, and user information.
Really Simple Syndication (RSS)

- RSS allows publishers to syndicate their content automatically.
- RSS data consists of summarized text and metadata such as publishing date.
- It uses a standardized XML-format to publish the content and can be read using RSS readers.
- The user subscribes to a feed by entering into the reader the feed's URI.
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>rss Title</title>
    <description>example of an rss feed</description>
    <link>http://www.rssnews.com/main.html</link>
    <pubDate>Mon, 06 Sep 2009 16:45:00 +0000</pubDate>
    <item>
      <title>Title for pub</title>
      <description>Text summary</description>
      <link>http://www.cnn.com/</link>
      <pubDate>Mon, 21 May 2012 16:45:00 +0000</pubDate>
    </item>
  </channel>
</rss>
Modifications to Yioop!

Existing Search in Yioop!

Web Images Video

Query Results: (Calculated in 6.816133 seconds. Showing results 0 - 10 of 279311)

The Avengers: Super Bowl trailer: Watch extended version here! | Inside Movies | EW.com
.. The 30-second version of the new trailer for The Avengers that ran today on the Super Bowl gave us just
http://insidenmovies.ew.com/2012/02/05/the-avengers-trailer-super-bowl/ Rank: 21.69 Rel: 44.25 Prox: 16.00
Score 9.80 Cached Similar Inlinks IP:74.200.247.69.

The Avengers | Inside Movies | EW.com
Movie Reviews See Categories See Archive RSS Category: The Avengers (1-10 of 14) Even more tantalizing:
Blink

Avenged Sevenfold Tab Pro | UG Plus @ Ultimate-Guitar.com
.. Avenged Sevenfold Tab Pro - View, learn and play along with Ultimate Guitar's huge database of high-quality
http://plus.ultimate-guitar.com/tab-pro/browse/a/avenged_sevenfold/ Rank: 10.75 Rel: 61.51 Prox: 11.50 Score

The Avengers: Super Bowl Trailer: Robert Downey Jr., Chris Evans, Scarlett Johansson And More Asse
If 14 seconds of "The Avengers" Super Bowl trailer wasn't enough to assemble, 30 seconds will do the trick!
The Super Bowl spot for "The Avengers" has arrived, and features all your favorite
http://news.moviefone.com/2012/02/05/the-avengers-super-bowl-trailer_n_1254999.html Rank: 11.21 Rel: 44.75

Joss Whedon on The Avengers' | Inside Movies | EW.com
according to writer-director Joss Whedon, "the filmmakers are a group of..."
With addition of social search

Query Results: (Calculated in 0.100391 seconds. Showing results 0 - 10 of 37)

Cobie Smulders introduces Marvel's The Avengers Video and Music Remix - YouTube
Cobie Smulders, who plays Maria Hill in Marvel's The Avengers, introduces the YouTube audience to Marvel's
http://www.youtube.com/watch?v=560LnGiKal&feature=g-feat Rank: 11.49 Ref: 26.60 Prox: 8.00 Score 9.95 Cached Similar Inlinks IP:74.125.224.07
http://www.tmz.com/member/me-1226/
... May Avengers 3D ...
http://www.tmz.com/member/me-1226/text] ... May Avengers 3D [jref=http://www.tmz.com/2012/05/02/jordan-hill-turns-himself-in-lakers/ Rank: 8.54 Ref: 18.52 Prox: 7.00 Score 9.76
http://dictionary.reference.com/browse/avengence
Avengence
http://ia.media-

Search results for keyword ‘avengers’
Social Search Architecture

- Search query
- Index
- Search Controller
- Web Crawl index
- Phrase Model
- Feed Model
- Feed Database
- View
- Web + Social search results
Feed Crawl Architecture

- Feed Server
- Twitter feed crawler
- RSS feed crawler
- Feed Database
- Twitter API
- RSS content providers

Flow:
- Request/Response
- Download XML files
Additions to Yioop!

- **Feed server**
  - Responsible for controlling the feed crawl activity.
  - To be started and stopped from command line
  - Initiates the RSS and Twitter feed crawler.

- **Feed model**
  - Responsible for the database operations of the search and administration activities of Yioop!
  - Responsible for the ranking of the search results

- **Feed Crawlers**
  - Twitter feed crawler
    - Queries the access tokens from database
    - Creates requests to Twitter API using access tokens
    - Parses the JSON data retrieved and inserts feed into database.
Additions to Yioop! Cont..

- RSS feed crawler
  - queries the feed URIs from the database
  - calls the Yioop! library curl function download pages
  - parses the XML content retrieved and inserts feed into database.

- Manage Feeds
  - Activity in the administration page for managing feed accounts.
  - Provides interface to add, delete RSS and Twitter feed subscriptions.

- Changes done in the search controller, search view, config file to accommodate social search results into Yioop!
Tables are added to the database to store the social data and user tokens.

- USER_KEYS
- FEED
- RSSFEED
- USER_RSSFEED
- USER_FEED

<table>
<thead>
<tr>
<th>FEED_ID</th>
<th>FEEDER</th>
<th>FEEDTEXT</th>
<th>REFEEDEDCOUNT</th>
<th>FEEDTIME</th>
<th>FEEDSOURCE</th>
<th>FEEDERPIC</th>
<th>FOLLOWERS_COUNT</th>
<th>FRIENDS_COUNT</th>
<th>VERIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>4633125815451650</td>
<td>OMGFacts</td>
<td>Due to pollution and overfishing</td>
<td>104</td>
<td>1337623313</td>
<td>&lt;a href=&quot;<a href="http://www.http://a0.twimg.con">http://www.http://a0.twimg.con</a>&quot;</td>
<td></td>
<td>4199143</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>463146502066689</td>
<td>BerkeleyHaas</td>
<td>Updates from IBD: Senegal [3 c</td>
<td>0</td>
<td>1337622917</td>
<td>&lt;a href=&quot;<a href="http://www.http://a0.twimg.con">http://www.http://a0.twimg.con</a>&quot;</td>
<td></td>
<td>10775</td>
<td>1931</td>
<td>0</td>
</tr>
<tr>
<td>4631464686522368</td>
<td>BerkeleyHaas</td>
<td>A whiteboard style discussion by</td>
<td>0</td>
<td>1337622917</td>
<td>&lt;a href=&quot;<a href="http://www.http://a0.twimg.con">http://www.http://a0.twimg.con</a>&quot;</td>
<td></td>
<td>10775</td>
<td>1931</td>
<td>0</td>
</tr>
<tr>
<td>4630741932445696</td>
<td>HarvardBiz</td>
<td>The Billion-Dollar Social Media Ql</td>
<td>76</td>
<td>1337622745</td>
<td>&lt;a href=&quot;<a href="http://twi">http://twi</a> <a href="http://a0.twimg.con">http://a0.twimg.con</a>&quot;</td>
<td></td>
<td>741086</td>
<td>177</td>
<td>0</td>
</tr>
<tr>
<td>4630737436147712</td>
<td>HarvardBiz</td>
<td>Marketing Needs a New Metaphor</td>
<td>53</td>
<td>1337622744</td>
<td>&lt;a href=&quot;<a href="http://twi">http://twi</a> <a href="http://a0.twimg.con">http://a0.twimg.con</a>&quot;</td>
<td></td>
<td>741086</td>
<td>177</td>
<td>0</td>
</tr>
<tr>
<td>46301930531249537</td>
<td>BarackObama</td>
<td>Mitt Romney’s firm bought Ampe</td>
<td>1295</td>
<td>1337622614</td>
<td>web</td>
<td></td>
<td>15706326</td>
<td>677979</td>
<td>1</td>
</tr>
</tbody>
</table>
Reverse Reciprocal Rank Fusion

\[
RRF_{\text{score}}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)},
\]

where

- \( d \) is document belonging to the set of documents \( D \).
- \( r \) is the ranking within the set \( R \).
- \( k \) is a constant

Ranking is based on metadata such as recency, retweets, followers count, friends count, verified account, etc.
Test and Results

Search results for the queries in different categories
(Tested on corpus of 3740 feeds)
Test and Results

- Average relevancy of the results among different categories

![Average relevancy among different categories](image-url)
Social search can be combined with web search for better search experience.

Social search is helpful in topics such as Movies, News, food etc.

Most of the social search results are relevant.
Questions

Thank You