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!

Project Goal

- 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.

Yioop! requirements

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 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 authors 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

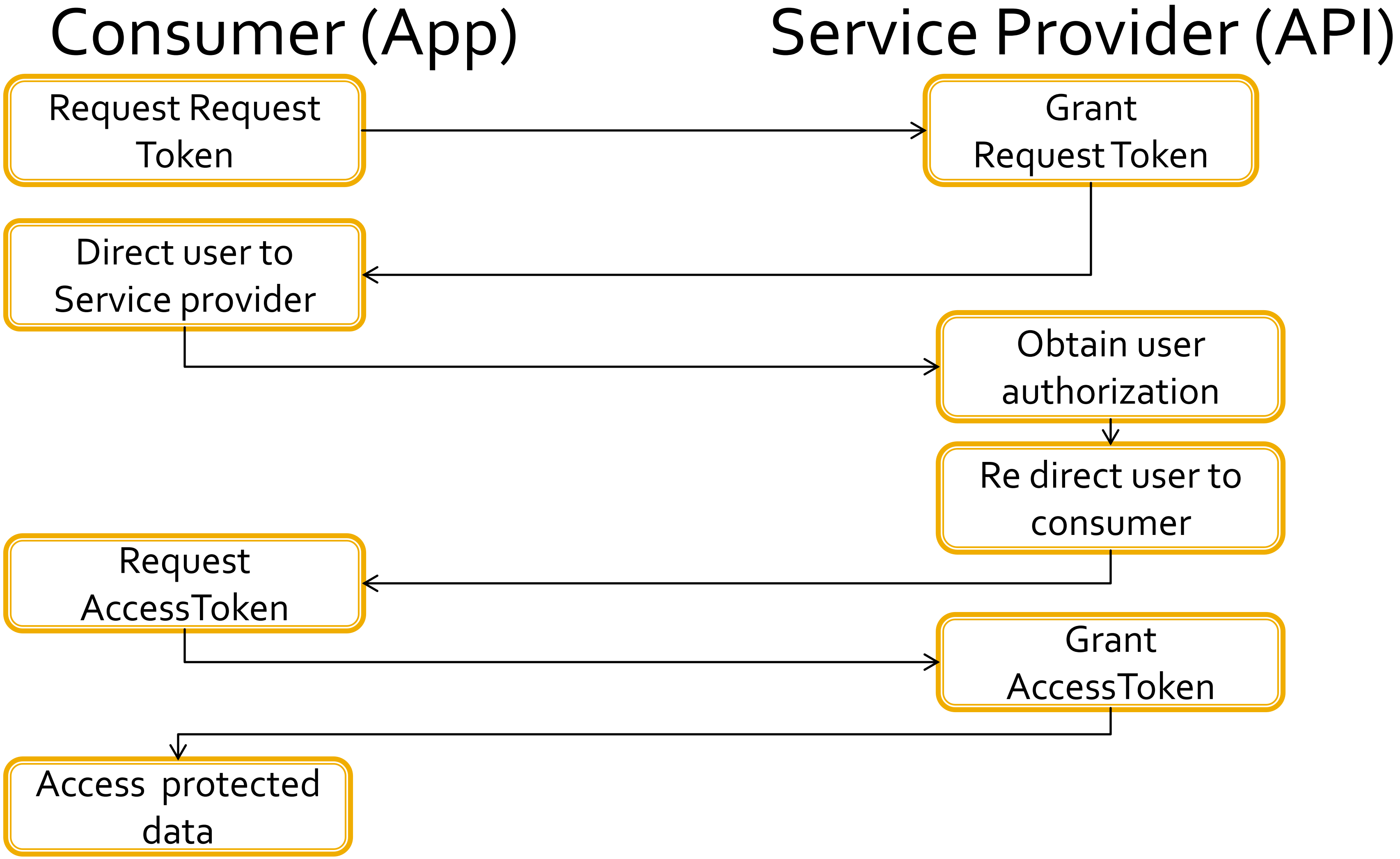
Twitter

- Twitter user’s content can accessed by twitter applications(consumers) authorized by user.
- Application can access content by making calls to APIs(providers). Example: REST API
- Twitter uses OAuth for authentication
- Each application has it’s own unique consumer key and consumer secret.

OAuth

- 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)

Service Provider (API)

Request Request Token

Grant Request Token

Direct user to Service provider

Obtain user authorization

Re direct user to consumer

Request AccessToken

Grant AccessToken

Access protected data

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:
-
```
[ {
  "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",
        ...
  },
 {
        ......
 },]
```

Really Simple Syndication(RSS)

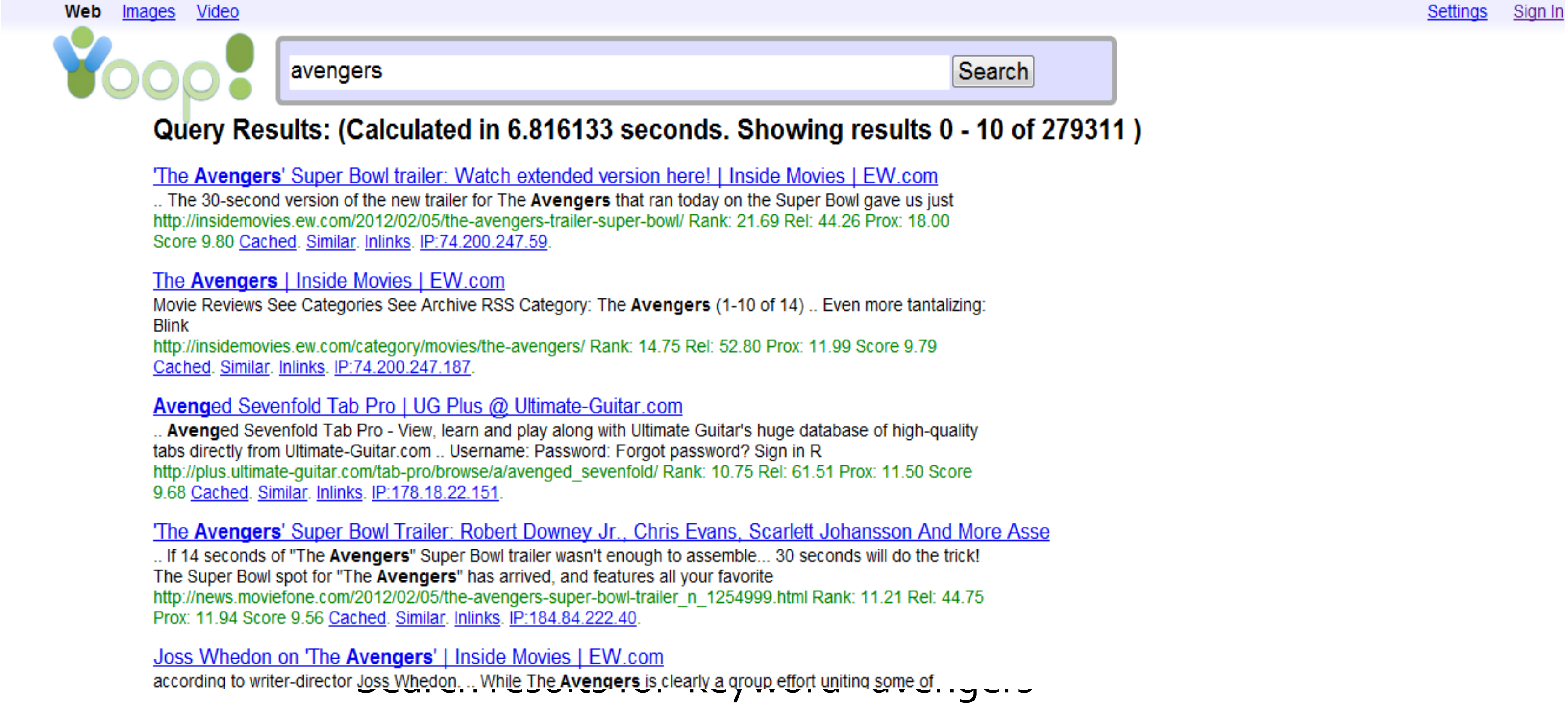
- RSS allows publishers to syndicate their content automatically.
- RSS data consists of summarized text and metadata such as publishing date.
- It uses a standardized XML-format to publish the content and can be read using RSS readers.
- The user subscribes to a feed by entering into the reader the feed's URI

Structure of RSS feed

```
<?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!



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://insidemovies.ew.com/2012/02/05/the-avengers-trailer-super-bowl/ Rank: 21.69 Rel: 44.26 Prox: 18.00 Score 9.80 Cached. Similar. Inlinks. IP:74.200.247.59.

The **Avengers** | Inside Movies | EW.com
Movie Reviews See Categories See Archive RSS Category: The **Avengers** (1-10 of 14) .. Even more tantalizing: Blink
http://insidemovies.ew.com/category/movies/the-avengers/ Rank: 14.75 Rel: 52.80 Prox: 11.99 Score 9.79 Cached. Similar. Inlinks. IP:74.200.247.187.

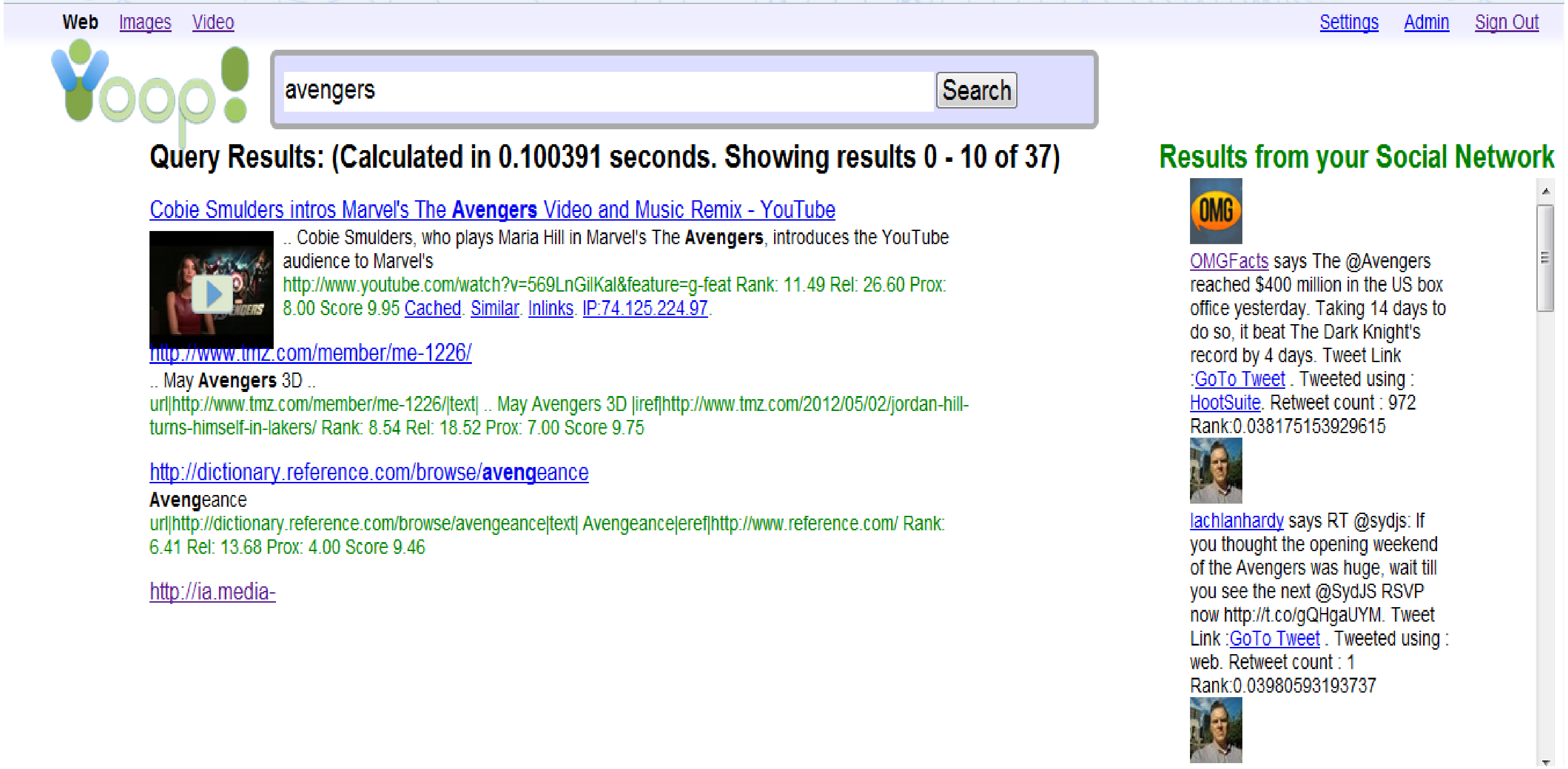
Avenged Sevenfold Tab Pro | UG Plus @ Ultimate-Guitar.com
.. **Aveng**ed Sevenfold Tab Pro - View, learn and play along with Ultimate Guitar's huge database of high-quality tabs directly from Ultimate-Guitar.com .. Username: Password: Forgot password? Sign in R
http://plus.ultimate-guitar.com/tab-pro/browse/a/avenged_sevenfold/ Rank: 10.75 Rel: 61.51 Prox: 11.50 Score 9.68 Cached. Similar. Inlinks. IP:178.18.22.151.

'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 Prox: 11.94 Score 9.56 Cached. Similar. Inlinks. IP:184.84.222.40.

Joss Whedon on 'The **Avengers**' | Inside Movies | EW.com
according to writer-director Joss Whedon. .. While The **Avengers** is clearly a group effort uniting some of

Search results for keyword "avengers"

With addition of social search



avengers

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

Cobie Smulders intros 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=569LnGilKal&feature=g-feat Rank: 11.49 Rel: 26.60 Prox: 8.00 Score 9.95 Cached. Similar. Inlinks. IP:74.125.224.97.

http://www.tmz.com/member/me-1226/

.. May **Avengers** 3D ..
url|http://www.tmz.com/member/me-1226/|text| .. May Avengers 3D |iref|http://www.tmz.com/2012/05/02/jordan-hill-turns-himself-in-lakers/ Rank: 8.54 Rel: 18.52 Prox: 7.00 Score 9.75

http://dictionary.reference.com/browse/**aveng**eance

Avengeance
url|http://dictionary.reference.com/browse/avengeance|text| Avengeance|eref|http://www.reference.com/ Rank: 6.41 Rel: 13.68 Prox: 4.00 Score 9.46

http://ia.media-

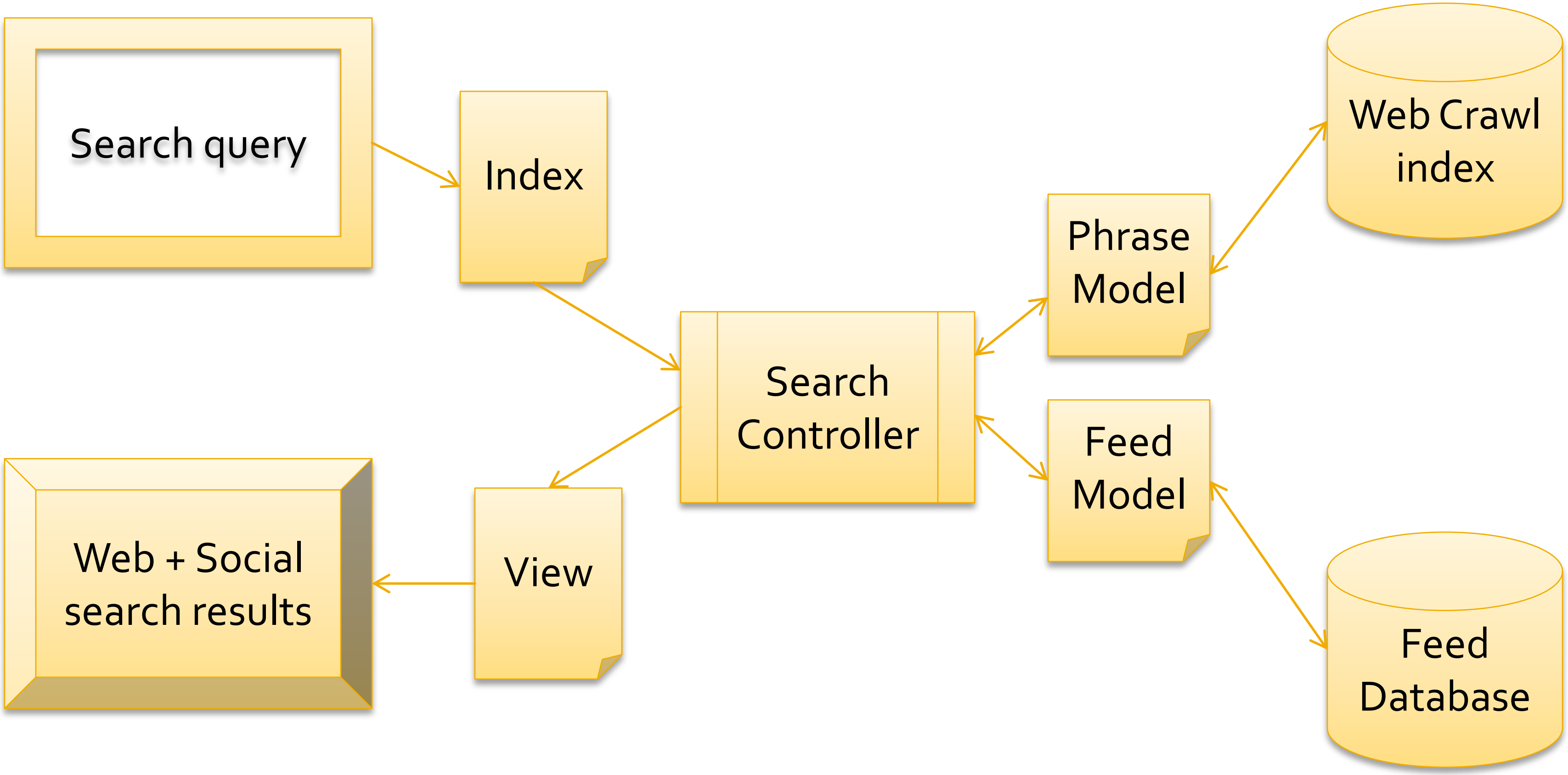
Results from your Social Network

OMGFacts says The @Avengers reached $400 million in the US box office yesterday. Taking 14 days to do so, it beat The Dark Knight's record by 4 days. Tweet Link :GoTo Tweet . Tweeted using : HootSuite. Retweet count : 972 Rank:0.038175153929615

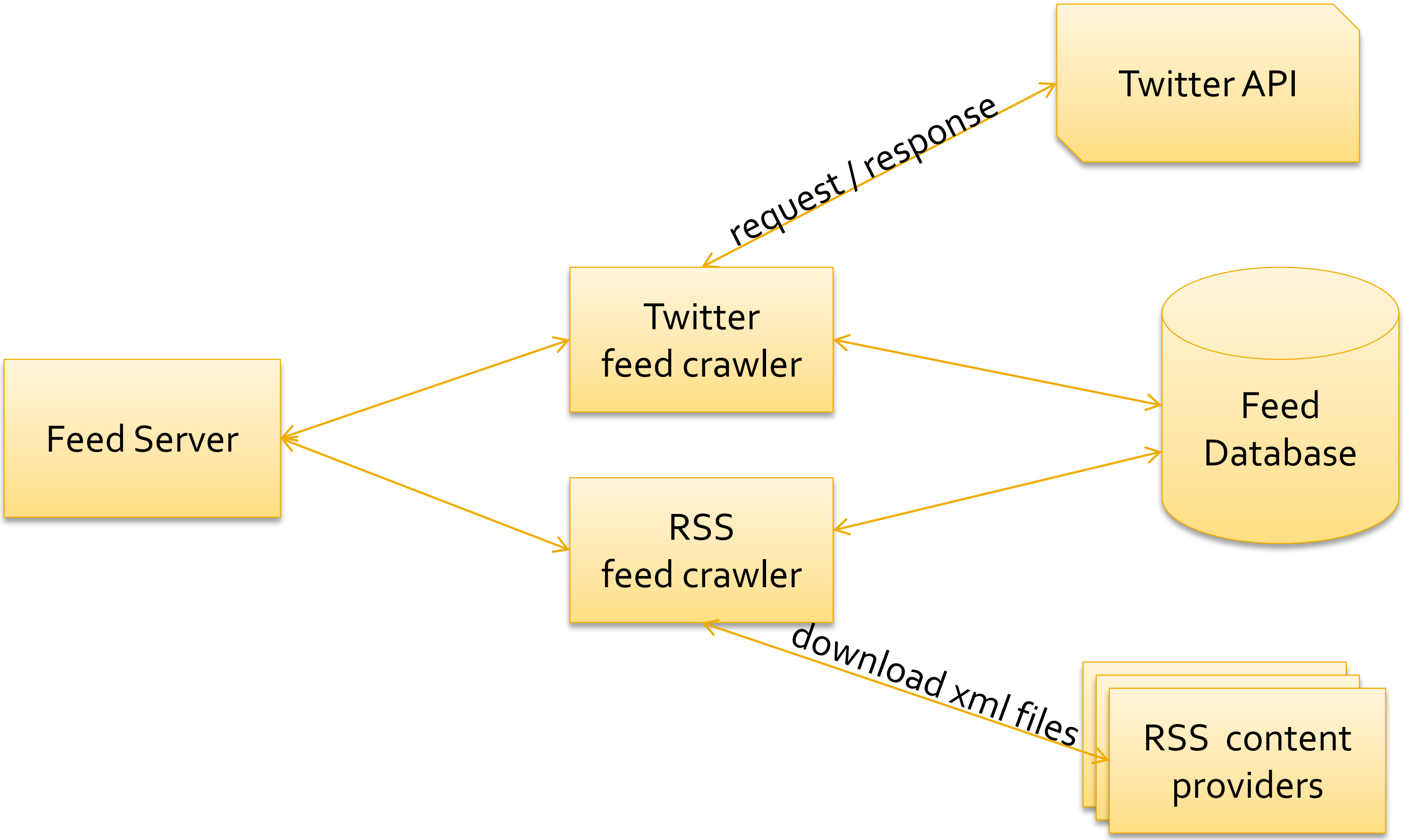
lachlanhardy says RT @sydjs: If you thought the opening weekend of the Avengers was huge, wait till you see the next @SydJS RSVP now http://t.co/gQHgaUYM. Tweet Link :GoTo Tweet . Tweeted using : web. Retweet count : 1 Rank:0.03980593193737

Search results for keyword 'avengers'

Social Search Architecture



Feed Crawl Architecture



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..

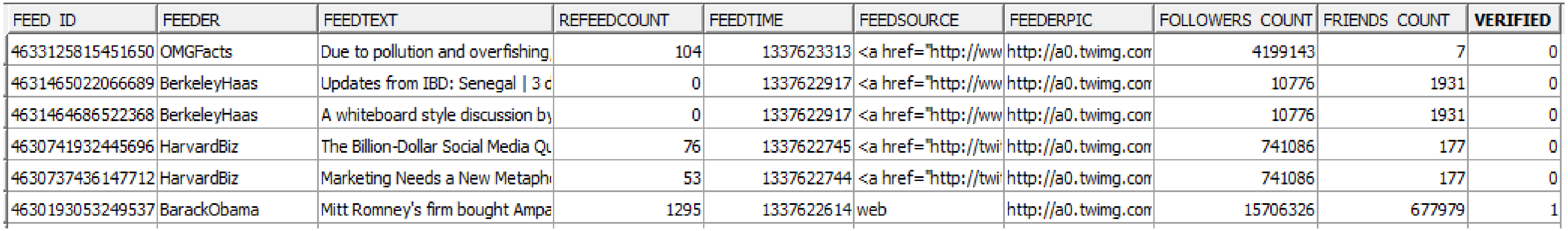
  - RSS feed crawler
    - queries the feed URIs from the database
    - calls the Yioop! library curl function download pages
    - parses the XML content retrieved and inserts feed into database.

- 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!

Additions to Yioop! Cont..

- Tables are added to the database to store the social data and user tokens.
  - USER_KEYS
  - FEED
  - RSSFEED
  - USER_RSSFEED
  - USER_FEED



FEED ID	FEEDER	FEEDTEXT	REFEEDCOUNT	FEEDTIME	FEEDSOURCE	FEEDERPIC	FOLLOWERS COUNT	FRIENDS COUNT	**VERIFIED**
4633125815451650	OMGFacts	Due to pollution and overfishing,	104	1337623313	<a href="http://ww	http://a0.twimg.com	4199143	7	0
4631465022066689	BerkeleyHaas	Updates from IBD: Senegal | 3 d	0	1337622917	<a href="http://ww	http://a0.twimg.com	10776	1931	0
4631464686522368	BerkeleyHaas	A whiteboard style discussion by	0	1337622917	<a href="http://ww	http://a0.twimg.com	10776	1931	0
4630741932445696	HarvardBiz	The Billion-Dollar Social Media Qu	76	1337622745	<a href="http://twi	http://a0.twimg.com	741086	177	0
4630737436147712	HarvardBiz	Marketing Needs a New Metaph	53	1337622744	<a href="http://twi	http://a0.twimg.com	741086	177	0
4630193053249537	BarackObama	Mitt Romney's firm bought Ampa	1295	1337622614	web	http://a0.twimg.com	15706326	677979	1

FEED table

Ranking function

- Reverse Reciprocal Rank Fusion

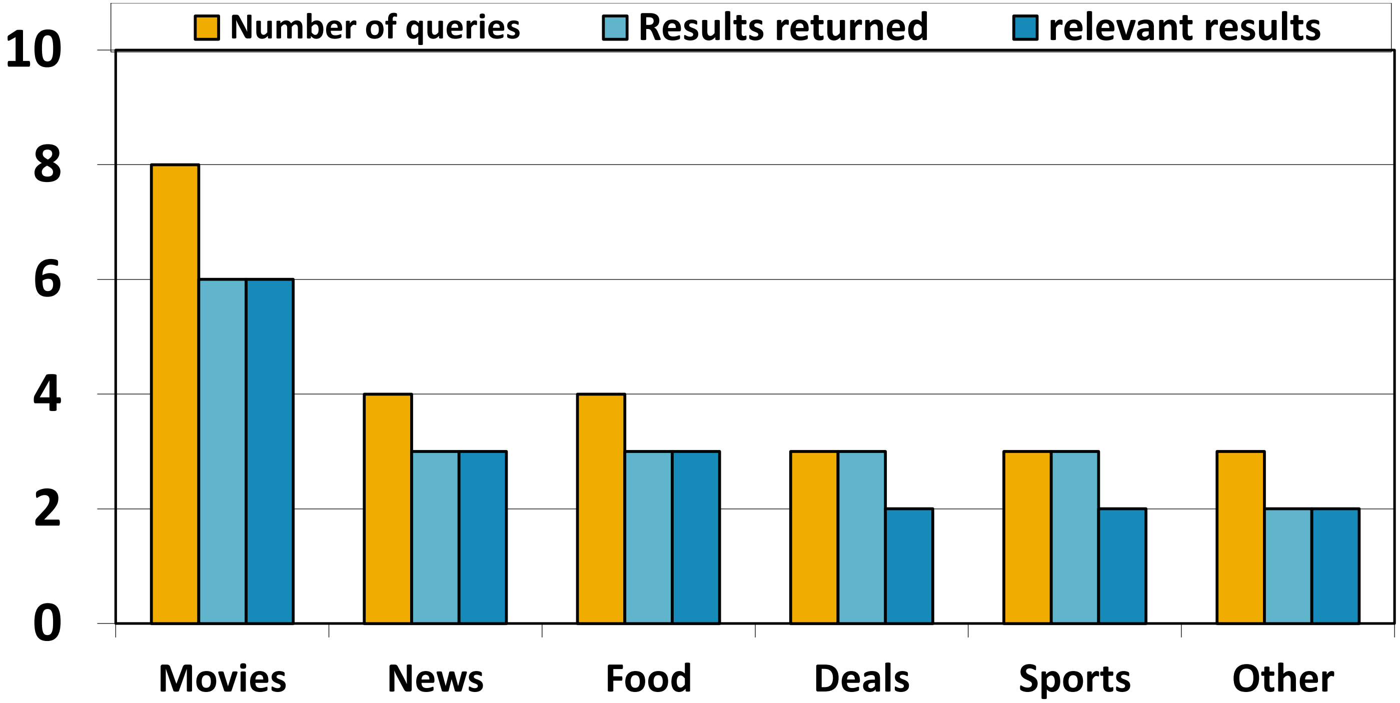
$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} ,$$ where

- ***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

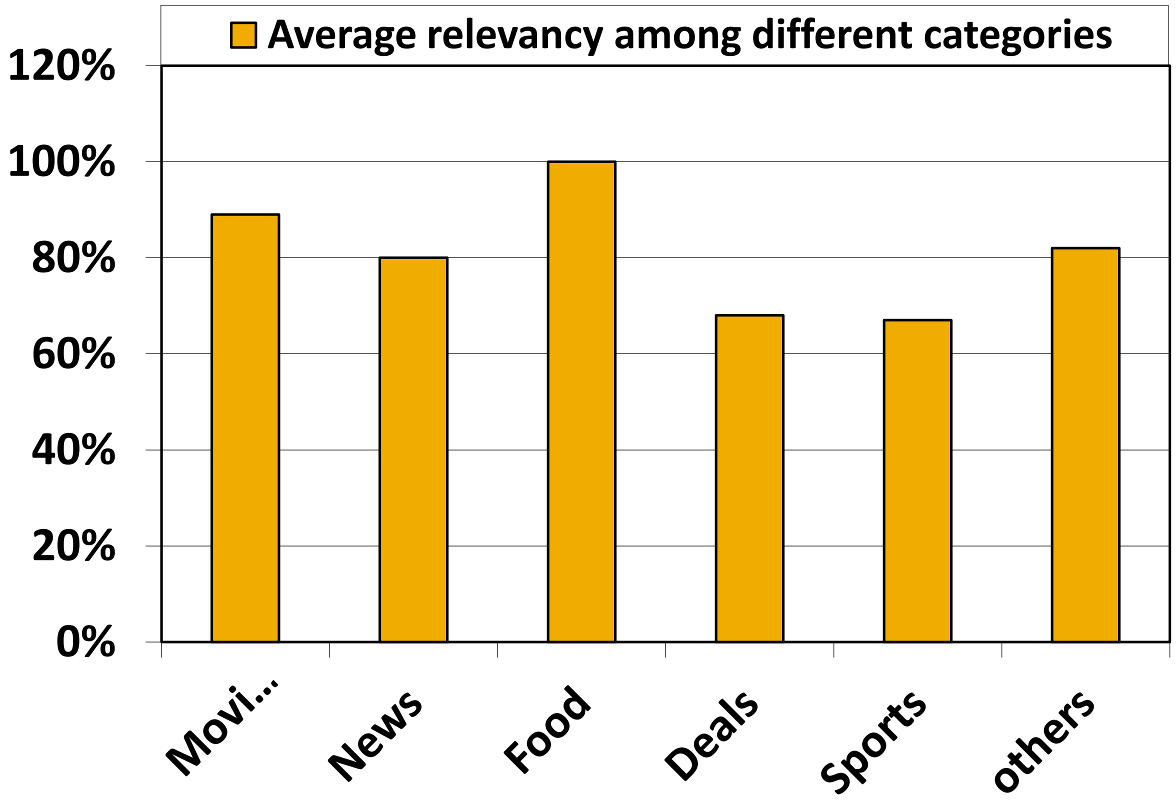
Test and Results

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



Conclusion

- 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