Yokie - Explorations in Curated Real-time Search & Discovery using Twitter

Owen Phelan, Kevin McCarthy, and Barry Smyth
CLARITY Centre for Sensor Web Technologies
School of Computer Science and Informatics
University College Dublin, Ireland
firstname.lastname@ucd.ie

ABSTRACT

Our research involves developing technology and techniques that apply the vast sea of real-time web data to interesting problems and topics. In this demo, we will present the ongoing development of a novel real-time search and discovery service named Yokie[1] (`http://yok.ie`, early technology description originally published in [1]). Yokie uses the large volume of hyperlink-laden messages on social networks like Twitter as the basis of its content and ranking systems. Curated sets of users (or "Search Parties") form the basis of sourcing the content from the networks, and the metadata of the containing messages form the basis of ranking and contextual retrieval of the hyperlinks. Each hyperlink is indexed with a compound set of terms from multiple tweets (should the given hyperlink be shared more than once). This indexing step is a novel example of collaborative tagging of resources. The application is live with more than 100 users, who have performed approximately 1000 queries. We will demonstrate the main techniques and novel ranking and retrieval techniques and user features.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Applications, Algorithms, Experimentation, Theory

Keywords

Search, Discovery, Information Retrieval, Relevance, Reputation, Twitter

1. DEMONSTRATION DESCRIPTION

The Yokie system can be broken down into several main components, all of which are briefly discussed in this section.

Curated "Search Parties"
A key aspect of the system is the curated list of content sources, what we've termed a *Search Party*. Users can curate dedicated search engines for personal and community use. These parties can be simple user lists, keywords, geographical metadata, or based on algorithmically generated lists. In the current prototype we have curated several seed Search Parties: *Technology*, *Irish Interest* and *World News* Twitter users (each containing 1000 users).

Content Acquisition
The system uses Twitter's API to acquire content as defined by the search parties. These messages are then stored and indexed using the service described in the following subsection. This component also carries out real-time language classification and finds other messages that contain the same URL so the system can calculate item popularity, and extracts tags so as to allow collaborative tagging.

Storage & Indexing
Once content is acquired, it is pushed to the Storage and Indexing subsystem. This is responsible for extracting metadata regarding the tweets, for instance timestamp data, hashtags (#linsanity, etc.), user profile information, location, etc. as well as the message content itself. The main content, the urlID of the URL mentioned in the message, and the timestamp is pushed to an indexer for storage and querying. In our current implementation we use Apache Solr[2] for this. We also store the remaining extracted metadata in a MongoDB[3] NoSQL database for easy retrieval and fast MapReduce functionality.

[1]Yokie - http://yok.ie

RecSys'12, September 9–13, 2012, Dublin, Ireland.
ACM 978-1-4503-1270-7/12/09.

[2]Apache Solr - http://lucene.apache.org/solr/
[3]MongoDB - http://mongodb.org

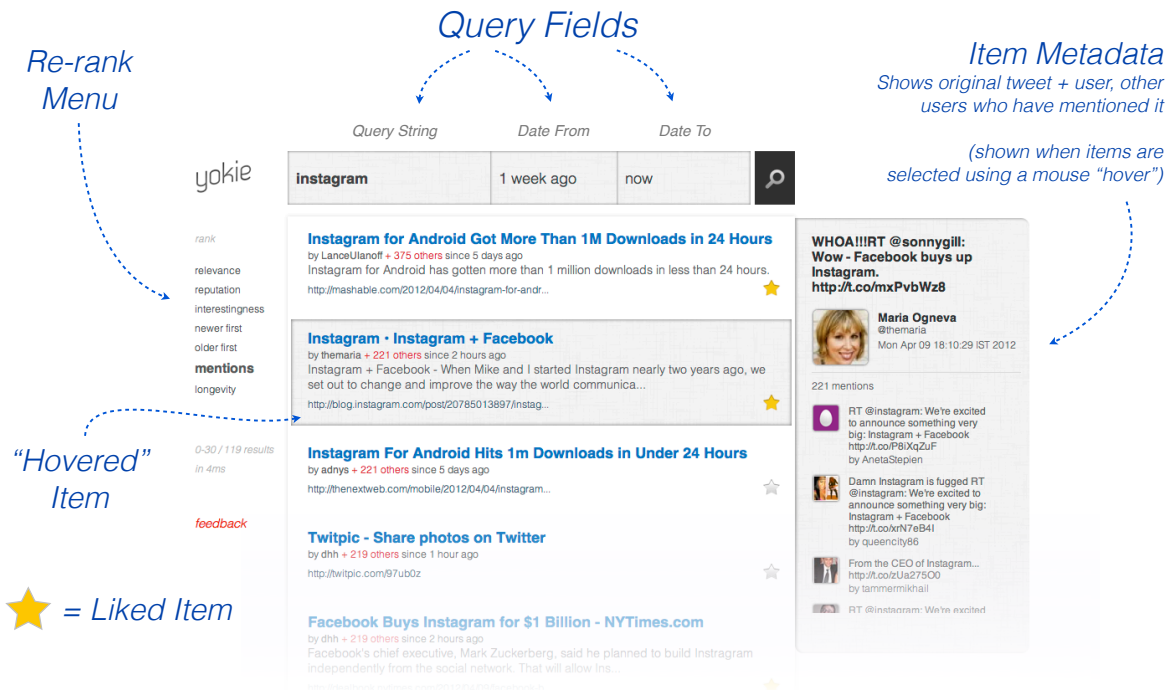

Figure 1: Yokie takes a traditional approach to Search-system layout, with query, ranking and results-list views. It also includes a UI pane for viewing extra metadata related to the item.

UI & Querying

The system is presented with a query interface, currently comprising of a query string field, and two temporal fields, *date from* and *date to*. The system takes as input a query string with an associated time window, which can be either a natural language query (e.g.. "1 day ago", "now", "last week", etc) or a fixed date ("12 June 2012"). These natural-language date strings are then parsed into UNIX timestamps. The UI also allows users to drill-down on results to explore related content such as the original tweet that the URL was shared with, the time and date it was shared, and the related Tweet mentions (if any). Each item contains a rating *star*, which users select when they believe an item is relevant to the query. The re-ranking menu allows users to re-rank the results. The querying subsystem parses user queries, based on the notation {QueryString, SearchParty, T_{max}, T_{min}}.

User & Item-based Result Ranking

Users can rank using typical relevance, which is vanilla Term Frequency Inverse Document Frequency scoring (TFxIDF) [2]. However, a set of ranking strategies beyond standard relevance have been devised using the added contextual metadata of the microblogs. These include ranking using temporal features such as item age and longevity (which is the size of the time window between the first and last mentions of the given hyperlink). Popular items can be ranked as we capture sample tweets that mention the hyperlink. Reputation is a user-based sum of follower-counts for each search-party member who have mentioned the link.

The premise of this strategy is high-volume in-links for users can be a sign of high user reputation. It is easily possible to derive a range of Reputation scores as we are specifically "listening" to people for content, as opposed to relying on a document graph that ignores content publishers. The novel Interestingness ranking algorithm is a function of popularity, level of user interactions of the items, and binary ratings.

2. CONCLUSION & FUTURE POTENTIAL

Yokie allows users to discover and rerank topical and niche webpages before they appear in old-style search engines like Microsoft's Bing. It has a considerable future in research of novel UI and ranking, collaborative search party curation, collaborative tagging, indexing and retrieval, user reputation, to name but a few.

3. ACKNOWLEDGEMENTS

This work is generously supported by Science Foundation Ireland under Grant No. 07/CE/11147 CLARITY CSET.

4. REFERENCES

[1] Owen Phelan, Kevin McCarthy, and Barry Smyth. Yokie - a curated, real-time search and discovery system using twitter. In *2nd Workshop of Real-time and Social Web. At Recommender Systems 2011*, RSWEB'11, 2011.

[2] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, March 2002.