An API-based Search System for One Click Access to Information

Dan Ionita
University of Twente
Enschede, The Netherlands

Niek Tax
University of Twente
Enschede, The Netherlands

Djoerd Hiemstra
University of Twente
Enschede, The Netherlands

Introduction

Next-generation search engines should be able to directly address user's information needs more often by providing the right information on the result page for certain types of information needs. Chilton et al [1] showed the current one click search implementation in Bing search to be regularly used, but also shows not every user's intent is always addressed in current one click systems.

This gave birth to our **research question**: *What is the best approach to design a One Click Access Information Retrieval system?*

For one-click search engines, the two most important features are being able to correctly identify the query's information need and using it to select the most relevant information sources. Web API's, if selected properly, can provide consistent, parse-able responses which can easily be used to gain information about the query. Such information from multiple specialized API's might then be used to classify the query as well as to construct a response.

In order to get insights in the best performing approaches we conducted a small literature study, focusing on the 1CLICK@NTCIR-9 participants. Based on findings in literature we began implemented a prototype one click system.

Approach

We implement a 3-tier, API-based approach:

1. **Retrieve features** by parsing query and analyzing results returned by various *API*s;
2. **Classify** into pre-defined categories based on retrieved features;
3. **Extract information** from various *API*s depending on category;

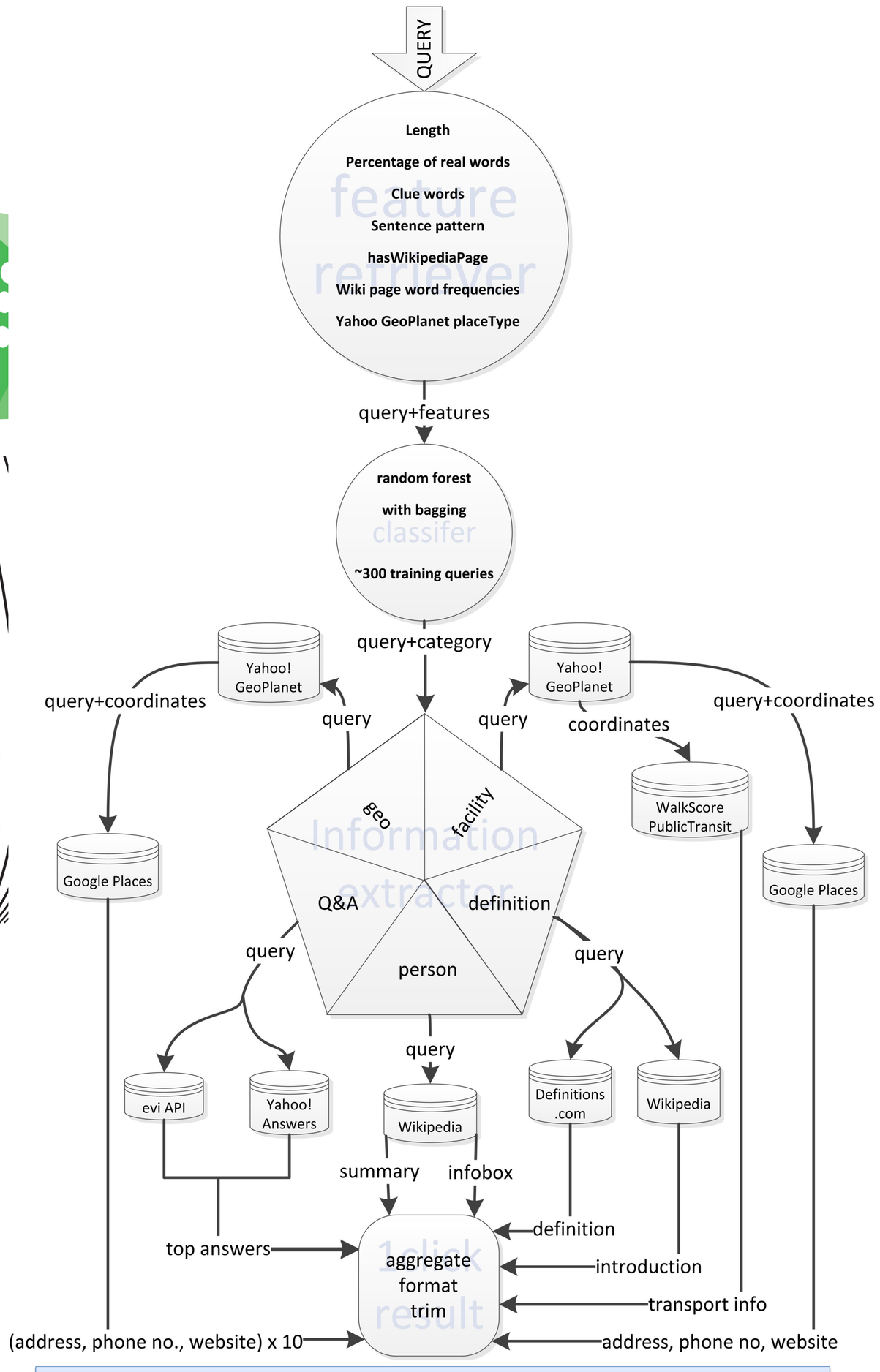

Fig. 1 *Data flow for query processing and output generation*

Design and fine-tuning

Manually created a set of queries used to for initial training. Baseline user study was then conducted with two-fold purpose: (1) get an impression of the *performance*, (2) gather more *training queries*.

Resulting set of *240 queries* (30 of each query category) used to evaluated the performance of various machine learning algorithms using the WEKA toolkit [2] and a stratified 10-fold cross validation. The Random Forest learner showed the highest classification accuracy, as can be seen in the confusion matrix below.

a	b	c	d	e	f	g	h	<– classified as
22	1	2	0	0	3	2	0	a = ARTIST
1	26	0	0	0	3	0	0	b = ACTOR
0	0	26	0	1	2	1	0	c = POLITICIAN
3	0	0	25	0	1	1	0	d = ATHLETE
2	0	1	0	16	7	4	0	e = FACILITY
1	0	0	0	0	29	0	0	f = GEO
2	0	1	2	3	0	21	0	g = DEFINITION
0	0	0	0	0	6	0	24	h = QA

Table. 1 *Confusion matrix for Random Forest (with bagging) classifier*

Feature retriever initially based on 1CLICK@NTCIR9 participants' approaches. After user study, the *performance evaluation* was used to iteratively improve feature-set used for classification and APIs used for information extraction.

Results

The system was evaluated using the S#-measure metric as part of the 1CLICK@NTCIR evaluation (described in [3]). Figure 2 below maps the results for English-Desktop runs (character limit 1000).
Both UT runs (implementing slightly different feature retrievers) scored highest amongst participant runs.

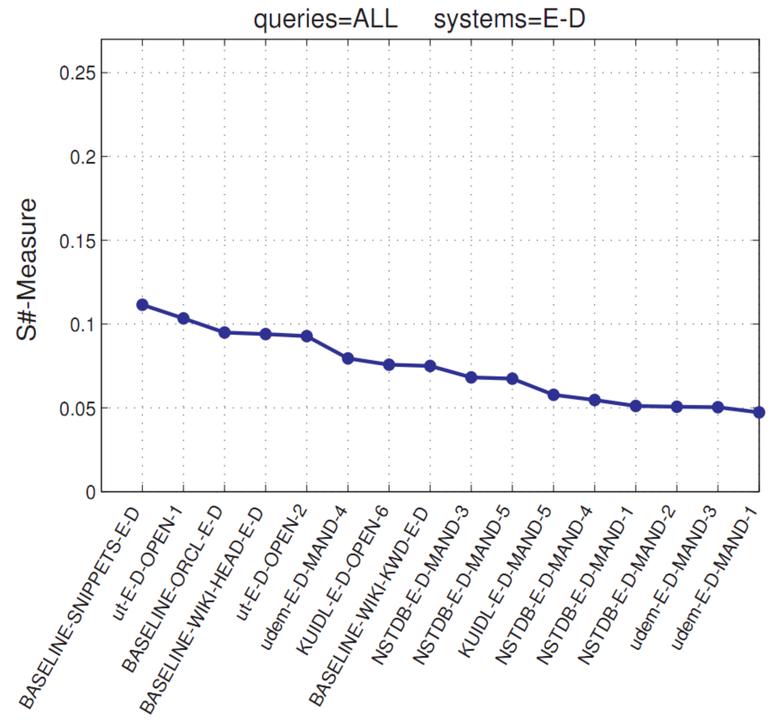

Fig. 2 *Ordered S# scores for all 1CLICK@NTCIR-10 Desktop runs*

References

1. L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*, WWW '11, 27-36, USA, 2011. ACM.
2. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10-18, 2009.
3. M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the ntcir-10 1click-2 task. In *Proceedings of the 10th NTCIR Conference*, 2013.

Acknowledgement

This publication was supported by the Dutch national program COMMIT and the Netherlands Organization for Scientific Research (NWO, project 639.022.809).

UNIVERSITY OF TWENTE.