Diversifying Search Results Using Time

Dhruv Gupta and Klaus Berberich

Max Planck Institute for Informatics

December 11, 2015

History-Oriented Queries

For our evaluation setup we considered three categories of history-oriented queries: categories of history-oriented queries : long-lasting wars, recurring events, and famous personalities. For constructing the queries we utilized reliable sources on the Web and data presented in prior research articles [2, 3]. We describe the details next.

Queries for long-lasting wars were constructed from the *WikiWars* corpus [3]. The corpus was created for the purpose of temporal information extraction. The keyword for the wars are given in Table 1a. For ambiguous important events we utilized the set of ambiguous queries used in the work by Gupta and Berberich [2]. The queries used are listed in Table 1b. For famous personalities we utilized a list of most influential people available on the USA Today [4] website. The names of these famous personalities were used based on the intuition that there would important events associated with them at different points of time. The list of all the entities is given in in Table 1c.

<table>
<tr><td><b>Americas</b> | american civil war | american revolution | mexican revolution</td></tr>
<tr><td><b>Europe</b> | world war II | world war I | french revolution | punic wars | spanish civil war | russo-polish war | second italo abyssinian war</td></tr>
<tr><td><b>Africa</b> | french algreian war | biafran nigerian civil war</td></tr>
<tr><td><b>Asia</b> | vietnam war | korean war | iraq war | persian wars | chinese civil war | iran iraq war | russian civil war | french indochina war | russo-japanese car</td></tr>
</table>

(a) Wars.

<table>
<tr><td><b>Sports</b> | commonwealth games | asian games | summer olympics | winter olympics | super bowl winners</td></tr>
<tr><td><b>Music</b> | u2 album | nirvana album | beatles album | red hot chilli peppers album | michael jackson album</td></tr>
<tr><td><b>Movies</b> | harry potter movie | oscar academy awards | lord of the ring movie</td></tr>
<tr><td><b>Politics</b> | german federal elections | us presidential elections | australia federal elections</td></tr>
</table>

(b) Events.

<table>
<tr><td><b>Business</b> | bill gates | sergey brin | larry page | howard schultz | sam walton</td></tr>
<tr><td><b>Science</b> | stephen hawking | francis collins | craig venter</td></tr>
<tr><td><b>Politics</b> | ronald reagan | mikhail gorbachev | george w. bush | deng xiaoping | nelson mandela | bill clinton | hillary clinton</td></tr>
<tr><td><b>Arts</b> | j. k. rowling | oprah winfrey | russell simmons | bono</td></tr>
<tr><td><b>Religion</b> | pope john paul II</td></tr>
<tr><td><b>Sports</b> | lance armstrong | michael jordan</td></tr>
<tr><td><b>Other</b> | ryan white | homer simpson | osama bin laden</td></tr>
</table>

(c) Entities.

Table 1: History-oriented queries.

In order to evaluate this diversified summary we obtain the corresponding *Wikipedia* [5] pages of the queries as ground truth summaries.

Description

Each *file* in our testbed is JSON file with the following schema :

```
 {
  "title": {
    "type": "string"
  },
  "publication_date": {
    "type": "string"
  },
  "queryKeywords": {
    "type": "string"
  },
  "goldSummary": {
    "type": "string"
  },
  "allTime": {
    "properties": {
      "charOffset": {
        "type": "string"
      },
      "isApprox": {
        "type": "boolean"
      },
      "range": {
        "type": "string"
      },
      "value": {
        "type": "string"
      },
      "word": {
        "type": "string"
      }
    }
  }
}
```

Figure 1: JSON schema for all the files in our testbed.

The title field denotes the title of the *Wikipedia* page. The publication_date field denotes the date on which the *Wikipedia* article was retrieved. The queryKeywords field denotes the keywords used as query for our system and baselines. The goldSummary denotes the actual body of the *Wikipedia* article corresponding to the title. Finally, allTime contains all the temporal expressions in the goldSummary; these were obtained by using SUTIME [1]. The sub-fields in allTime are properties of the various temporal expressions provided by SUTIME.

Citation

This data set was compiled for our publication at ECIR 2016:

> Gupta D. and Berberich K.
> *Diversifying Search Results Using Time*,
> ECIR 2016.

Kindly, cite our work when using the dataset in your research.

References

[1] Chang A. X. and Manning C. D. Sutime: A library for recognizing and normalizing time expressions. In *LREC 2012*, pages 3735–3740, 2012.

[2] Gupta D. and Berberich K. Identifying time intervals of interest to queries. In *CIKM 2014*, pages 1835–1838, 2014.

[3] P. P. Mazur and R. Dale. Wikiwars: A new corpus for research on temporal expressions. In *EMNLP 2010*, pages 913–922, 2010.

[4] USA Today. Top-25 influential people, 2015. [Online; accessed 23-September-2015] `http://usatoday30.usatoday.com/news/top25-influential.htm`.

[5] Wikipedia. Wikipedia, the free encyclopedia, 2015. [Online; accessed 23-September-2015] `http://en.wikipedia.org/`.