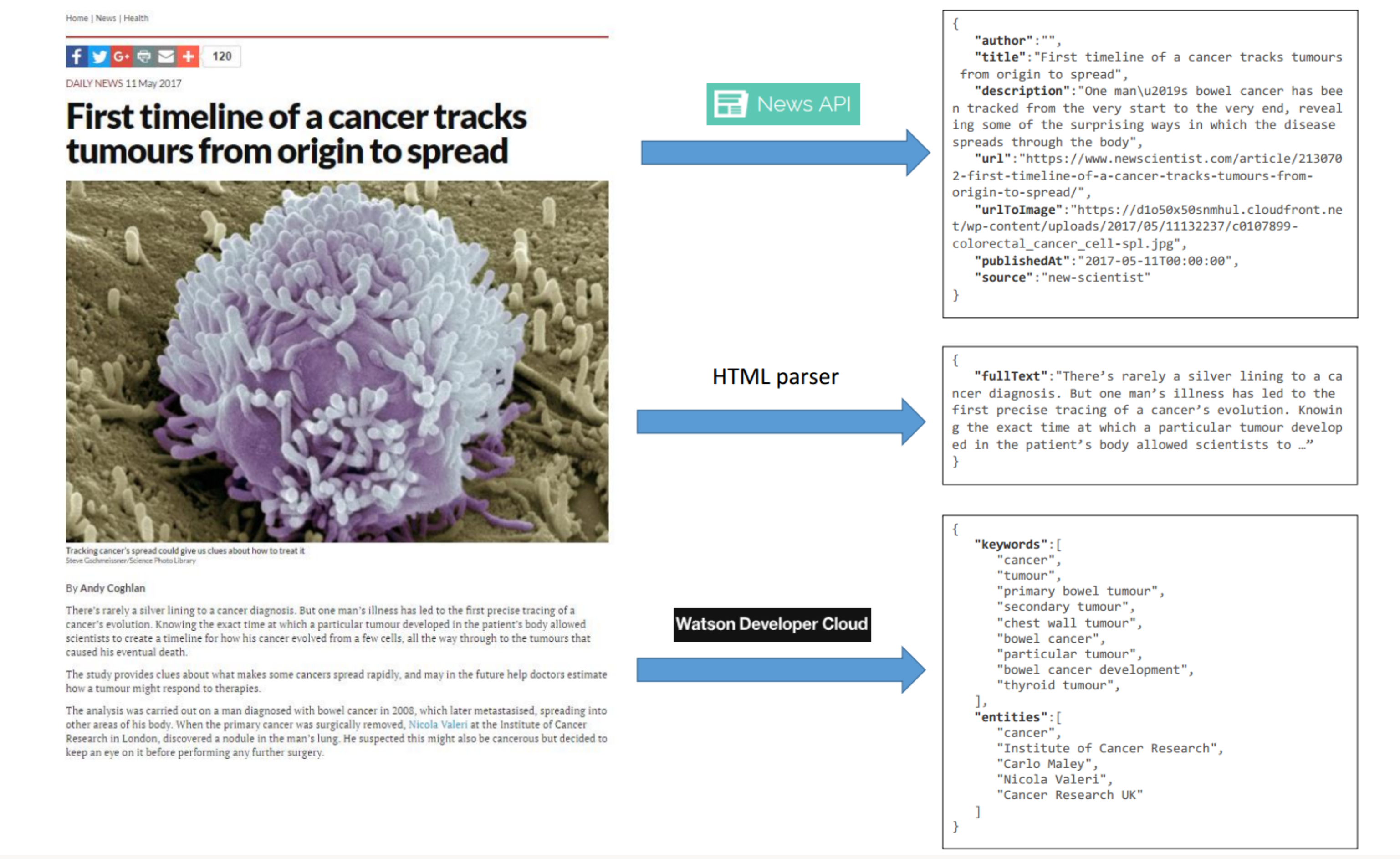
News Recommendation System

Introduction

In today's society a news user is often overwhelmed by the amount of available articles, spread over thousands of different topics. Reading the front page of a news site, will most likely result in articles that we do not find relevant and explicitly searching for a topic is often an effort that we do not want to make. News recommendation services therefore aim at providing articles that the user wants to read, without the need to search for them.

In this work we explores the effectiveness of integrating a user's reading history in the recommendation engine, in the form of articles that the user has liked or visited. Furthermore, we consider different sources of meta-data that can better describe the content of an article for the purpose of producing recommendations.

The information content of an article



Home | News | Health

120

DAILY NEWS 11 May 2017

First timeline of a cancer tracks tumours from origin to spread

Tracking cancer's spread could give us clues about how to treat it

Steve Gschmeissner/Science Photo Library

By Andy Coghlan

There's rarely a silver lining to a cancer diagnosis. But one man's illness has led to the first precise tracing of a cancer's evolution. Knowing the exact time at which a particular tumour developed in the patient's body allowed scientists to create a timeline for how his cancer evolved from a few cells, all the way through to the tumours that caused his eventual death.

The study provides clues about what makes some cancers spread rapidly, and may in the future help doctors estimate how a tumour might respond to therapies.

The analysis was carried out on a man diagnosed with bowel cancer in 2008, which later metastasised, spreading into other areas of his body. When the primary cancer was surgically removed, Nicola Valeri at the Institute of Cancer Research in London, discovered a nodule in the man's lung. He suspected this might also be cancerous but decided to keep an eye on it before performing any further surgery.

```
{
   "author":"",
   "title":"First timeline of a cancer tracks tumours
 from origin to spread",
   "description":"One man\u2019s bowel cancer has bee
n tracked from the very start to the very end, reveal
ing some of the surprising ways in which the disease
spreads through the body",
   "url":"https://www.newscientist.com/article/213070
2-first-timeline-of-a-cancer-tracks-tumours-from-
origin-to-spread/",
   "urlToImage":"https://d1o50x50snmhul.cloudfront.ne
t/wp-content/uploads/2017/05/11132237/c0107899-
colorectal_cancer_cell-spl.jpg",
   "publishedAt":"2017-05-11T00:00:00",
   "source":"new-scientist"
}
```

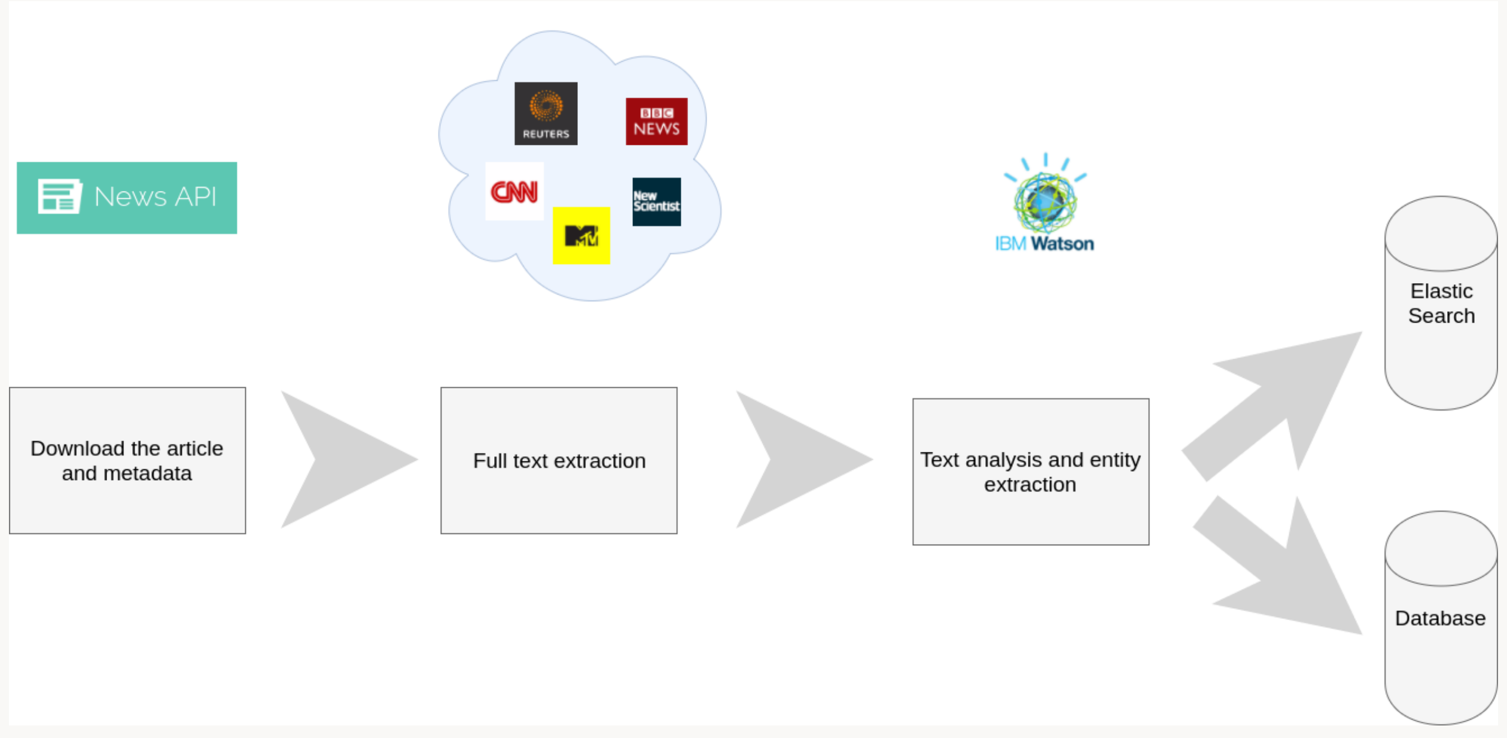
```
{
   "fullText":"There's rarely a silver lining to a ca
ncer diagnosis. But one man's illness has led to the
first precise tracing of a cancer's evolution. Knowin
g the exact time at which a particular tumour develop
ed in the patient's body allowed scientists to …"
}
```

```
{
   "keywords":[
      "cancer",
      "tumour",
      "primary bowel tumour",
      "secondary tumour",
      "chest wall tumour",
      "bowel cancer",
      "particular tumour",
      "bowel cancer development",
      "thyroid tumour",
   ],
   "entities":[
      "cancer",
      "Institute of Cancer Research",
      "Carlo Maley",
      "Nicola Valeri",
      "Cancer Research UK"
   ]
}
```

Figure: Example of metadata extraction from a news

- To be able to identify the content of a piece of news we do not rely uniquely on its title of text, but we leverage several sources of meta-data.
- Simple meta-data can be the author, published date and publishing source
- More advanced (and useful) characteristics can be found extracting entities and keywords through natural language processing techniques
- Enriching the documents in storage with these information is a key step in our processes

Data ingestion pipeline



News API
IBM Watson
Elastic Search
Download the article and metadata
Full text extraction
Text analysis and entity extraction
Database

Figure: Overview of the pipeline

- First we retrieve news from the *News API* service, The API acts as an aggregator for different news websites and additionally returns meta-data such as author, publishing date, source and a short description
- After that, the full text of every article is scraped from the source using a web crawler
- Then we use the natural language processing services from *IBM Watson Developer Cloud* to extract keywords and entities from the full text
- Finally, we persist the documents in a database and we index them in an Elasticsearch instance

Evaluation

In our evaluation we track the Discounted cumulative gain (DCG) scores assigned by the users to our recommendations as a function of the information we hold on their preferences.

1. Create N_Q queries to be used in the news search engine
2. Every participant sequentially execute these queries and evaluates the recommendations:

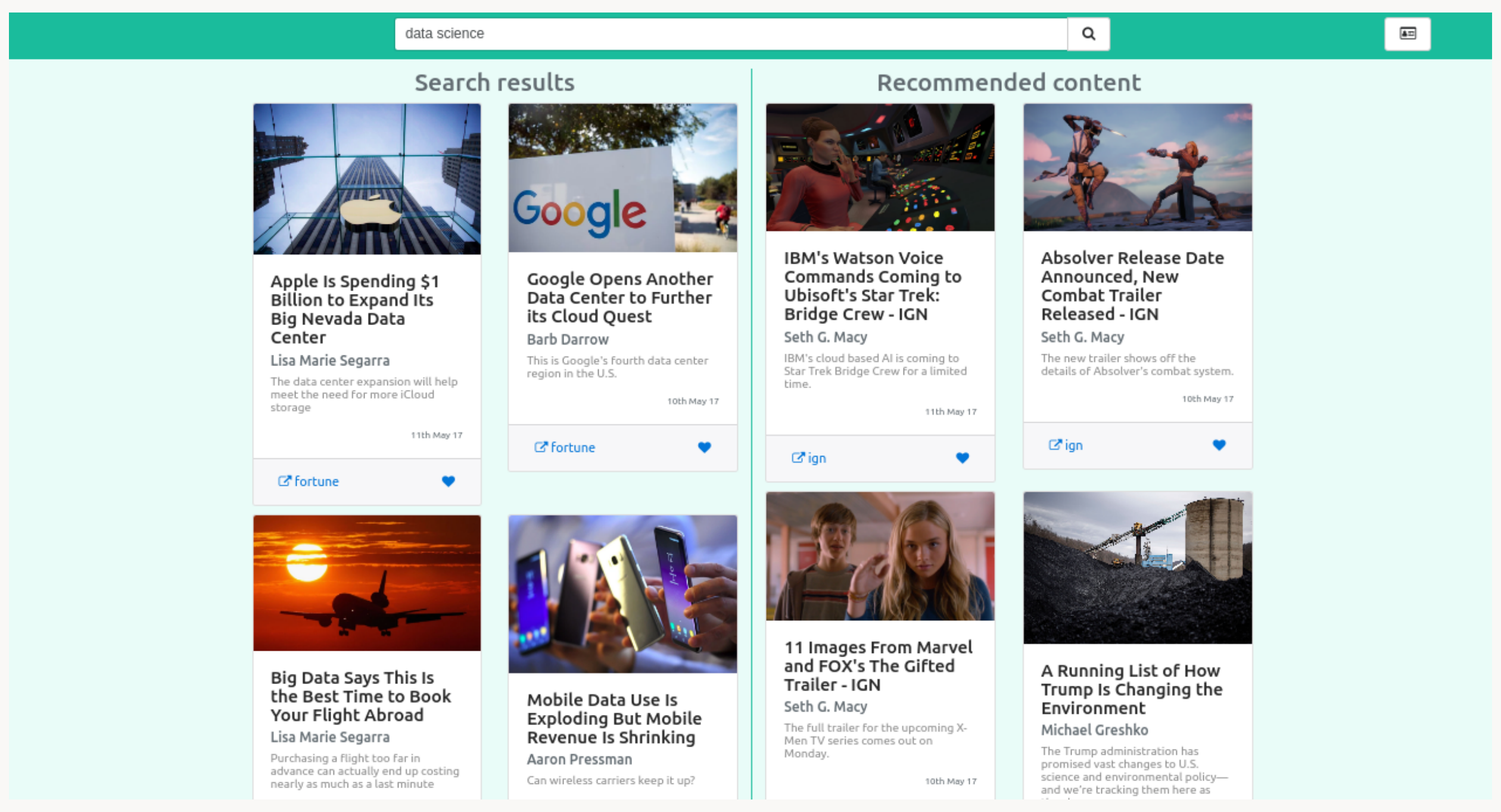
   2.1 Run a query in the news article search engine

   2.2 Consider the list of articles retrieved and choose the first 2 documents that he or she finds interesting starting from the top

   2.3 Look at the $K = 5$ articles that appear in the recommendation list and assign them a relevance score based on the aforementioned table

   2.4 Run another randomly selected query

User interface



data science

Search results

Apple Is Spending $1 Billion to Expand Its Big Nevada Data Center
Lisa Marie Segarra
The data center expansion will help meet the need for more iCloud storage
11th May 17
fortune

Google Opens Another Data Center to Further its Cloud Quest
Barb Darrow
This is Google's fourth data center region in the U.S.
10th May 17
fortune

Big Data Says This Is the Best Time to Book Your Flight Abroad
Lisa Marie Segarra
Purchasing a flight too far in advance can actually end up costing nearly as much as a last minute

Mobile Data Use Is Exploding But Mobile Revenue Is Shrinking
Aaron Pressman
Can wireless carriers keep it up?

Recommended content

IBM's Watson Voice Commands Coming to Ubisoft's Star Trek: Bridge Crew - IGN
Seth G. Macy
IBM's cloud based AI is coming to Star Trek Bridge Crew for a limited time.
11th May 17
ign

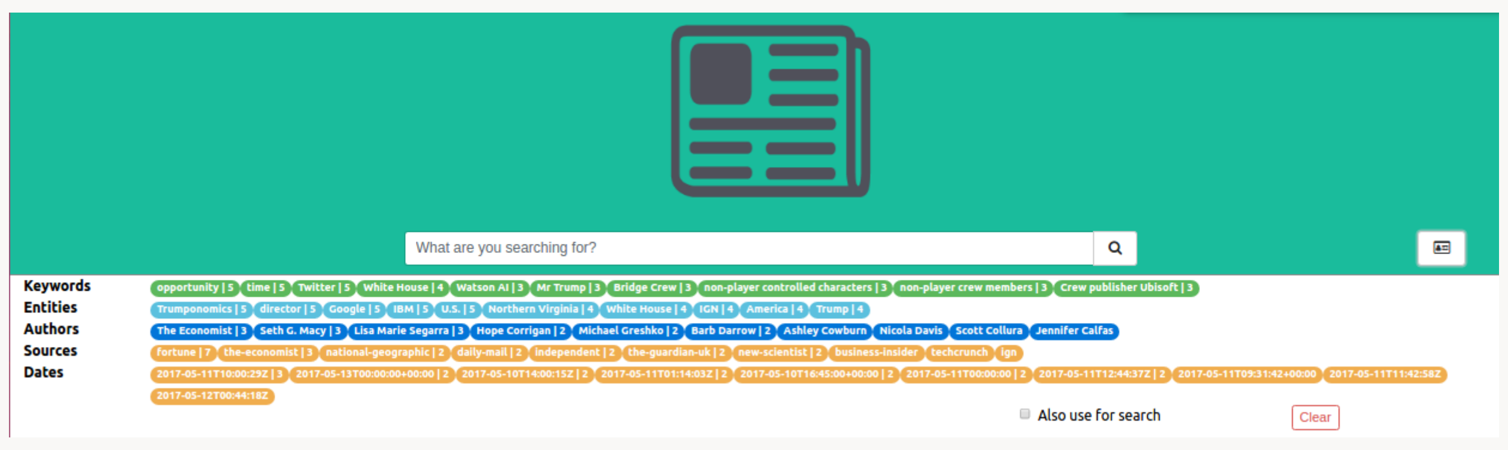
Absolver Release Date Announced, New Combat Trailer Released - IGN
Seth G. Macy
The new trailer shows off the details of Absolver's combat system.
10th May 17
ign

11 Images From Marvel and FOX's The Gifted Trailer - IGN
Seth G. Macy
The full trailer for the upcoming X-Men TV series comes out on Monday.
10th May 17

A Running List of How Trump Is Changing the Environment
Michael Greshko
The Trump administration has promised vast changes to U.S. science and environmental policy—and we're tracking them here as

Figure: Front-end application.

- The user can search for queries and see the results (left side), like in a normal search engine
- The user is given recommendation (right side), these are continously updated.



What are you searching for?

Keywords: opportunity | 5, time | 5, Twitter | 5, White House | 4, Watson AI | 3, Mr Trump | 3, Bridge Crew | 3, non-player controlled characters | 3, non-player crew members | 3, Crew publisher Ubisoft | 3

Entities: Trumponomics | 5, director | 5, Google | 5, IBM | 5, U.S. | 5, Northern Virginia | 4, White House | 4, IGN | 4, America | 4, Trump | 4

Authors: The Economist | 3, Seth G. Macy | 3, Lisa Marie Segarra | 3, Hope Corrigan | 2, Michael Greshko | 2, Barb Darrow | 2, Ashley Cowburn, Nicola Davis, Scott Collura, Jennifer Calfas

Sources: fortune | 7, the-economist | 3, national-geographic | 2, daily-mail | 2, independent | 2, the-guardian-uk | 2, new-scientist | 2, business-insider, techcrunch, ign

Dates: 2017-05-11T10:00:29Z | 3, 2017-05-13T00:00:00+00:00 | 2, 2017-05-10T14:00:15Z | 2, 2017-05-11T01:14:03Z | 2, 2017-05-10T16:45:00+00:00 | 2, 2017-05-11T00:00:00 | 2, 2017-05-11T12:44:37Z | 2, 2017-05-11T09:31:42+00:00, 2017-05-11T11:42:58Z, 2017-05-12T00:44:18Z

□ Also use for search

Clear

Figure: User's interest view.

- The user profile, or interest view, is based on the articles that the user has liked or visited from the search engine results
- It is possible to clear the user's profile

Results

- Discounted cumulative gain for a *top-K* recommendation system

$$DCG = \sum_{k=1}^{K} \frac{rel(k)}{\log_2(k+1)}$$

- The results of our user study:
  - 11 people in the test group / 11 people in the control group
  - 5 articles to rank in the recommendation list
  - 10 search queries in random order, allowing to select 2 results from each query

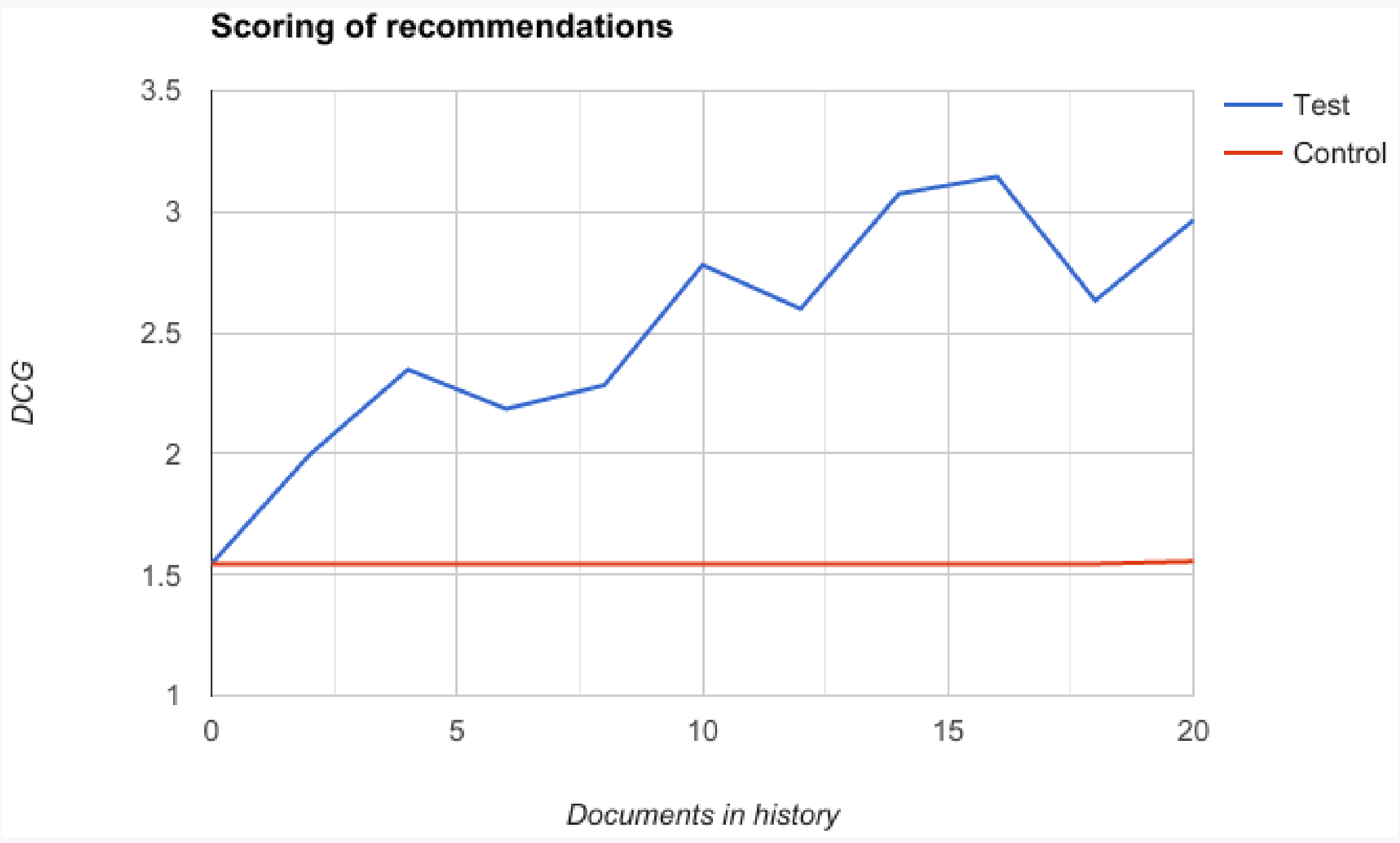


Figure: Evolution of DCG score for the test and control groups

- User profiling and customization are the base for every recommendation system
- We have validated the usage of meta-data as a key ingredient of producing recommendations
- The relative weight of every meta-data factor can be tweaked to vastly improve the results

Federico Baldassarre, Sandra Bäckström, Sebastian Bujwid, Zimin Chen

KTH Royal Institute of Technology