SIGNAL: Finding relevant news at scale

David Corney
Search Solutions 2015
The challenge

Collecting, analysing and distributing one million news articles every day in a continuous stream
Research at Signal Media

- 30 staff including 5 researchers
  - Machine learning, NLP, IR
- Regular academic visitors / partners
- Co-ordinating NewsIR workshop at ECIR2016
  - Release of 1m article collection for research
  - http://research.signalmedia.co/newsir16
Product overview

- Acquire c. 1 million news/blog posts per day
- Annotate people, organizations, topics, locations, sources...
- Store in ElasticSearch
- Feeds tailored to individual’s needs
- Clients include brand monitoring, PR, market intelligence etc.
Empire Global adds experience to board and management team

Casino City Times

24 November 2015 (PRESS RELEASE) – Empire Global Corp. has announced it has appointed Graham Martin to the management team and the board of directors of the company, increasing the size of the board to four members. Martin, based in London and Hong Kong, is a third generation retail bookmaker and widely recognised expert in the online gaming industry. Read original article

Also appears in 1 more source

Rolling Good Times

Yesterday at 4:37pm

Read article
Cameron rejects Corbyn's concerns over looming NHS 'winter crisis'

Prime minister jokes during PMQs that he would award Labour leader 'full Marx' for creating his own winter crisis in his party.
Where (near-)duplicates come from

Scenario 1: aggregation

Article appears on a news website
Copy later appears on news aggregation site
Where (near-)duplicates come from

Scenario 2: syndication

Copies of same article appear on multiple sites at the same time
(e.g. regional papers)
Where (near-)duplicates come from

Scenario 3: breaking news updates

Article appears on a news site
Same article is then updated on same site
Why duplicates cause problems

- Irrelevant results for most users
  - Just want to see “original” or “authoritative” version
  - C.f. plagiarism detection
- But some users want to see all content
  - PR companies measuring impact
Isn’t detecting duplicates easy?

- Exact string matching is simple
  - But sensitive to small changes
- Two articles are near-duplicates if their cosine similarity is close to one
  - Or use edit distance
- But pairwise comparison of 1m articles/day is hard work
Generate and refine candidates

- Quickly generate shortlist of candidate duplicates
- Pairwise analysis of shortlist is fast
- Can then select/rank candidates
- Scales up
Initial collection → Shortlist of candidates → Final results

http://signalmedia.co

@SignalHQ
Duplicate detection with hashing

● Generate hash code for every article
  ○ Fixed-length representation
  ○ Similar articles have similar codes (LSH)

● Shortlist: collisions with existing articles

● Refine: pairwise comparison (if needed)
Hashing example (simplified)

Article 1  →  n-grams  →  [Diagram of hash values]
Article 2  →  n-grams  →  [Diagram of hash values]
Article 3  →  n-grams  →  [Diagram of hash values]

Collision (duplicate)
Duplicate detection through search

- Articles already indexed in Elasticsearch
- Extract key features of article to form a query
  - Significant terms, named entities...
- Shortlist: use query to search for similar articles
- Refine: pairwise comparison (if needed)
Search example (simplified)

Jeremy Corbyn has warned of the risk of another "winter crisis" in the NHS as he and David Cameron clashed over health at Prime Minister's Questions.

● Extract key terms from article:
  ○ “Cameron” “winter” “NHS” “Corbyn”...

● Search for
  ○ (Cameron + winter + NHS) or
  ○ (Cameron + NHS + Corbyn) or
  ○ (NHS + winter + Corbyn) etc.
Practical issues and heuristics

- Identical titles may be sufficient to show duplication
- Regular distinct articles
  - E.g. diary pieces: same title, different content
- Parallel / concurrent writing of duplicates to collection
Evaluation

- Relative distance between pairs of articles
  - Levenshtein edit-distance per character
  - Too slow to make decisions, but useful for evaluation

- Mean distance between pairs of articles
  - labelled as duplicates: 24.5%
  - not labelled as duplicates: 80.5%

- Of those labelled as duplicates
  - 27% are identical
  - rest are edited versions of same text
One day at Signal

Out of c. 1 million articles received...

○ 500k have a title we’ve already seen
○ Another 100k have contents we’ve already seen

...leaving 400k distinct articles per day
Thanks for listening!

Contact

E: david.corney@signalmedia.co
T: @SignalHQ
W: signalmedia.co

News Information Retrieval Workshop

http://research.signalmedia.co/newsir16