Improving content discovery through combining linked data and data mining techniques

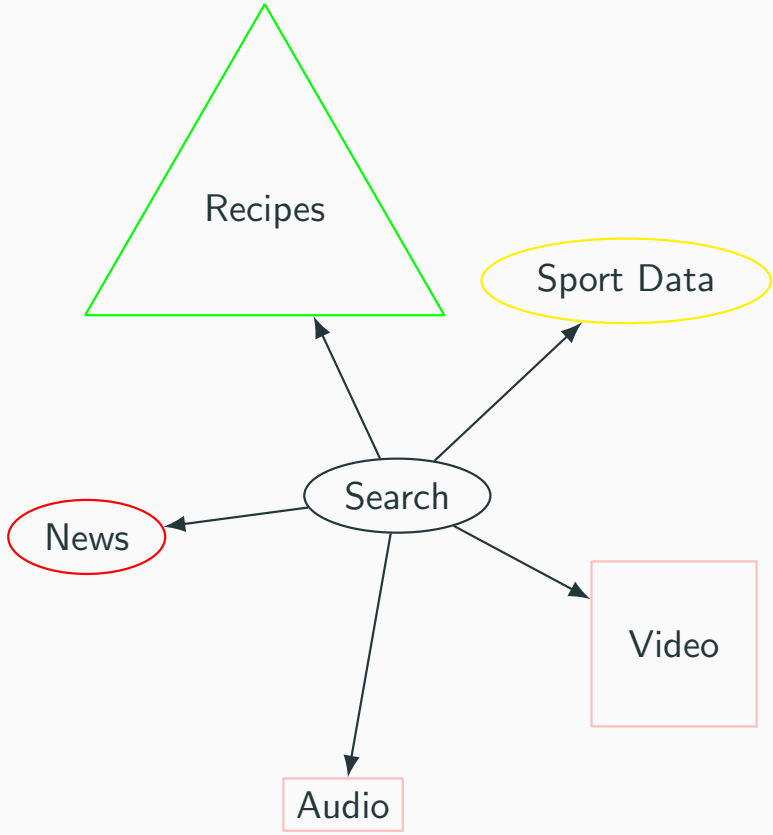
Ross Fenning, Principal Software Engineer

2016-05-16

BBC Design and Engineering: Content Discovery

Problem

- In the BBC, content is in multiple data stores
- >10,000,000s of content items
- Content is diverse, systems are incompatible, etc.

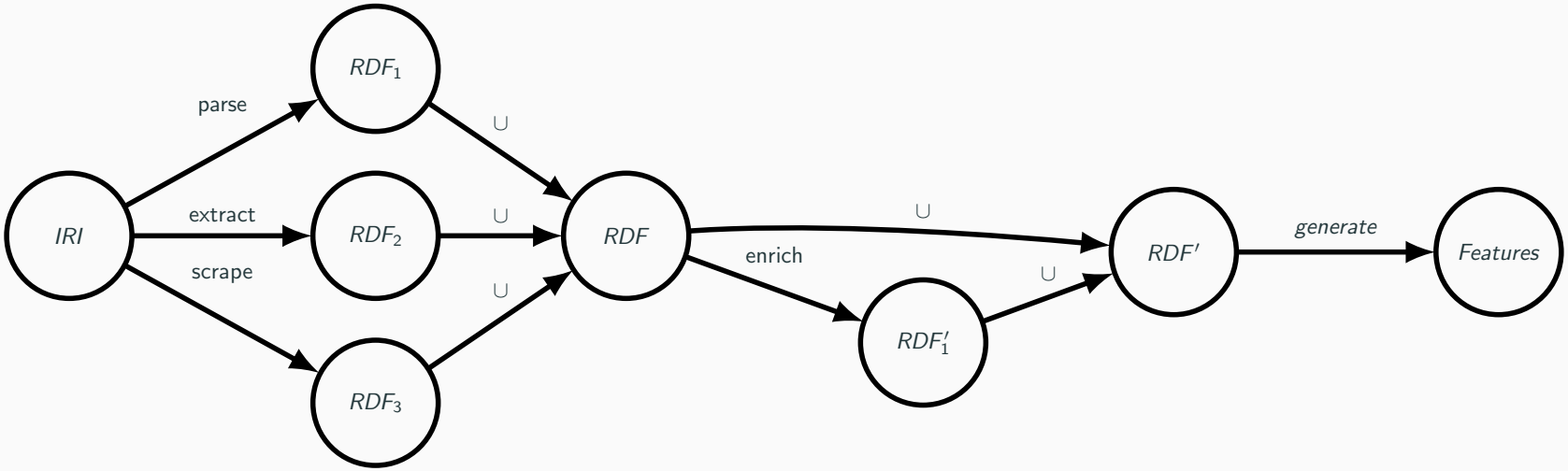


Cross-content functions like search, navigation, recommendations, etc. are difficult to implement.

How can we learn enough about all our content and offer people links to interesting parts of the website they've not seen before?

Approach

Data pipeline using semantics and linked data



- Construct RDF graphs from content items using linked data, entity extraction and more
- Enable or disable each and union each to get a single RDF graph
- Enrich that graph by extracting data about linked objects (other enrichment considered, but not implemented)
- Generate feature sets using SPARQL queries
- Apply clustering to feature sets

Outcomes

Eureka! How to make discoveries at the speed of light - BBC News

How cloud computing is enabling organisations to analyse vast swathes of data in real time and derive valuable insights faster than ever before.

Extracting RDFa, etc. from the page finds basic categoriation for social media embedding. This article is from BBC News (technology news).

Surveyed people preferred related content based on topics and themes. Entity extraction does this, but needs tuning.

The Roskilde music festival in Denmark is a huge logistical undertaking, attracting 130,000 visitors each year.
Over eight days guests camp out and watch 170 live performances. They buy and consume 200 tonnes of food and generate 300 tonnes of waste.
Running the event efficiently and safely is a must, and now a team at Copenhagen Business School (CBS) says it has found a way to help organisers do just that - using "cloud analytics".
Working with computing giant IBM, it began to collect mountains of real-time data during the

Embedded semantics are trivial to extract, so start here. Better clustering comes from combining with entity extraction, if tuned.

Encrichment using embedded semantics of linked entities adds little. Further research is needed for more sophisticated enrichment methods.