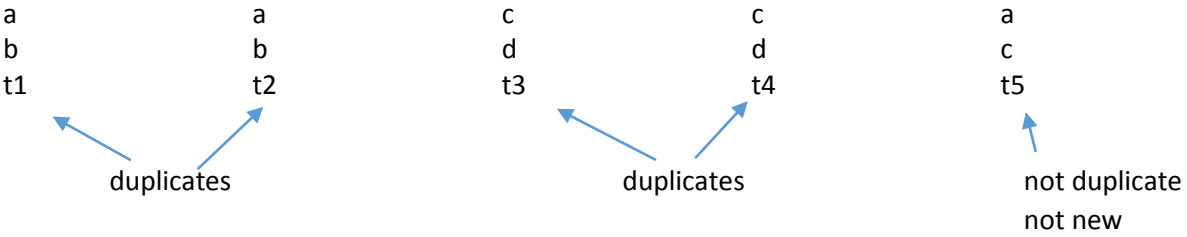
Note Taker: Neslihan Bulut

Week of April 14, 2014

**Novelty Detection in DDS(Document Data Stream)**

Find documents that contain new information.



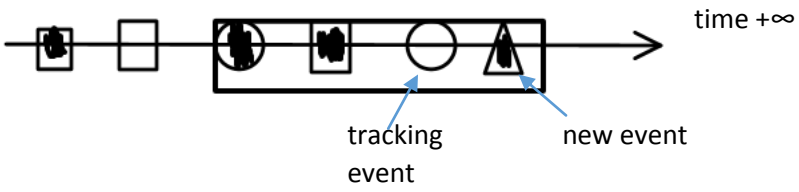

Conclusion: Not duplicate does not necessarily mean that it is new.

<u>New Event Detection</u>

1. The first document represent a new event
2. Compare the new article with the existing new events if different enough it indicates a new event.

Training -> Find parameters-> Use these parameters for testing

After some time old becomes new -> Time window




Information retrieval within the context of novelty implies diversity.

User submits an ad-hoc query.

Search engines tries to provide everything (as much as possible) about different meanings of the ad-hoc query.

Query = jaguar (cat / car / operating system / cocktail)

One possible approach is search result clustering.

Carrot

| | |
|---|---|
| $Cluster_1$Label | |
| $Cluster_2$Label | |

Query Types

Ambiguous queries: jaguar

Underspecified Queries: James Bond (music, movie, book) -> Auto completion of queries by search engines

Information gathering: how to cook pizza

Miscellaneous: Aimed to find a specific product

Providing answers to different meanings of the same query is referred to as search result diversification problem (SRD).

SRD Approaches

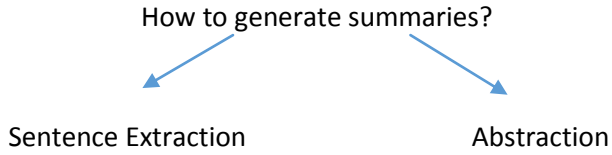
1. Implicit: Use the set of documents returned by the search engine for diversification (MMR).
2. Explicit: Use the query for diversification. Find different meanings of the query words and use them as separate queries and present them to the user.

MMR

Maximal marginal relevance.

The use of MMR diversity base ranking for reordering documents and producing summaries, SIGR 1998.

J. Carbonell, J. Goldstein (2 page article)




Query-based summarization

Find the sentence which is most similar to the query.

MMR = $argMax_{d_i \in R \setminus S} \left[ \lambda sim_1(d_i, q) - (1 - \lambda) max_{d_j \in S} sim_2(d_i, d_j) \right]$

C = Collection
$d_i, d_j \in C$

Relevant documents in C

R\S = R – S

S : Selected documents so far to be presented to the user.

$$\lambda = [0,1] \quad \lambda \uparrow higher\ accuracy\ and\ relevance, \qquad \lambda \downarrow higher\ diversity$$

$$S' = \begin{bmatrix} 1 & 0.11 & 0.23 & 0.76 & 0.25 \\ & 1 & 0.29 & 0.57 & 0.51 \\ & & 1 & 0.02 & 0.20 \\ & & & 1 & 0.33 \\ & & & & 1 \end{bmatrix}$$ Similarity among the top 5 documents selected by the search engine

Search engine sim(Q,d)

| d1 | d2 | d5 | d3 | d4 |
|---|---|---|---|---|
| 0.91 | 0.90 | 0.63 | 0.50 | 0.06 |

$\lambda$ =1 Rank according to relevance

d1, d2, d5, d3, d4

$\lambda$ =0 Similarity according to diversity

d1 sim(d1, d2) = 0.11 (-.11) → highest

sim(d1, d3) = 0.23 (-.23)

sim(d1, d4) = 0.76 (-.76)

sim(d1, d5) = 0.25 (-.25)

S = {d1, d2} R={d3, d4, d5}

sim(d1, d3) = 0.23 sim(d2, d3) = 0.29

sim(d1, d4) = 0.76 sim(d2, d4) = 0.57

sim(d1, d5) = 0.25 sim(d2, d5) = 0.51

d3 → S = {d1, d2, d3} ; selecting the least similar

$\lambda = 0.5$

1st iter: S={d1} without MMR

Use MMR the first time to select the next doc.

R\S = {d2, d3, d4, d5}

MMR(d2)= 0.5 * 0.90 – 0.5*0.11 = 0.395

MMR(d3) = 0.5 * 0.5 – 0.5 * 0.23 = 0.135

MMR(d4) = 0.5 * 0.06 – 0.5 *0.76 = -0.35

MMR(d5) = 0.5 * 0.63 – 0.5 * 0.25 = 0.19

Max MMR value is observed with d2 S={d1, d2}

2nd use of MMR S={d1, d2}, R\S={d3, d4, d5}

MMR(d3) = 0.5 * 0.5 – 0.5 * 0.29 = 0.105

sim(d1, d3) = 0.23 sim(d2, d3) = 0.29

d4: sim(d1, d4) = 0.74 sim(d2, d4) = 0.57

MMR(d4) = 0.5 * 0.06 – 0.5 * 0.74 = -0.34

d5: sim(d1, d5) = 0.25 sim(d2, d5) = 0.51

MMR(d5) = 0.5 * 0.63 – 0.5 * 0.51 = 0.06

Highest MMR is observed with d3.

So S becomes S= {d1, d2, d3}

Compare $\lambda = 1$ all based on relevance

S = {d1, d2, d5} S(d1, d2) = 0.11 + S(d1, d5) = 0.25 + S(d2, d5) = 0.51 = 0.87

$\lambda = 0, \lambda = 0.5$

S= {d1, d2, d3} sim(d1, d2) = 0.11 + sim(d1, d3) + sim(d2, d3) = 0.29 = 0.63

0.63 < 0.87 case 2 provides more diversity.

Total similarity(relevance) to the query

<u>Case 1</u>

$sim_1$(d1, q) = 0.91

$sim_1$(d2, q) = 0.90

$sim_1$(d5, q) = 0.63

sum of all similarities above: 2.44

<u>Case 2</u>

S = {d1, d2, d3}

$sim_1$(d1, q)=0.91

$sim_1$(d2, q)=0.90

$sim_1$(d3, q)=0.50

sum of all similarities above:2.3

2.44 > 2.3

2.44 has less diversity more relevance to the query.

**Exercises:**

1. Based on the similarity matrix and the query results given below, with the MMR criteria having $\tau = 1$ and $\tau = 0.5$ show the top 3 ranked documents;

$$S = \begin{pmatrix} 1 & 0.28 & 0 & 0 & 0.5 \\ & 1 & 0.33 & 0.57 & 0.28 \\ & & 1 & 0.66 & 0 \\ & & & 1 & 0.50 \\ & & & & 1 \end{pmatrix}$$

Search engine results;

| Document | Similarity |
|---|---|
| d1 | 0.07 |
| d2 | 0.90 |
| d3 | 0.60 |
| d4 | 0.76 |
| d5 | 0.03 |

**Answer**

$\tau = 1$ selects the documents based on their relevance to the query, hence the result is; d2, d4, d3.

$\tau = 0.5$ ; Start with S = {d2}  R-S={d1, d3, d4, d5}

MMR(d1) = 0.5 * 0.07 – 0.5 * 0.28 = -0.105

MMR(d3) = 0.5 * 0.60 – 0.5 * 0.33 = 0.135

MMR(d4) = 0.5 * 0.76 – 0.5 *0.57 = 0.095

MMR(d5) = 0.5 * 0.03 – 0.5 * 0.28 = -0.125

Highest MMR is observed with d3, hence S = {d2, d3}, R-S = { d1, d4, d5}

MMR(d1) = 0.5 * 0.07 – 05 * 0.28 = -0.105

MMR(d4) = 0.5 * 0.76 – 0.5 * 0.66 = 0.05

MMR(d5) = 0.5 * 0.03 – 0.5 * 0.28 = -0.125

Highest MMR is observed with MMR(d4), hence S becomes {d2, d3, d4}