Information Integration in the Domain of News Articles

Michal Holub
holub@fiit.stuba.sk

Supervisor: Mária Bieliková

Current state and hypothesis

- many news portals report about the same event
- the user has to search for more information about an event—no place integrates the articles

Hypothesis: News articles referring about the same event share at least 50 % of some specific feature.

Goals

- integrate information about an event
- group news articles from many sources
- present the news articles in an integrated way to the user

Domain specifics

- not every news portal informs about all events
- two news portals do not publish similar article in the same time
- we want to group an article as soon as it is published

Events of the day

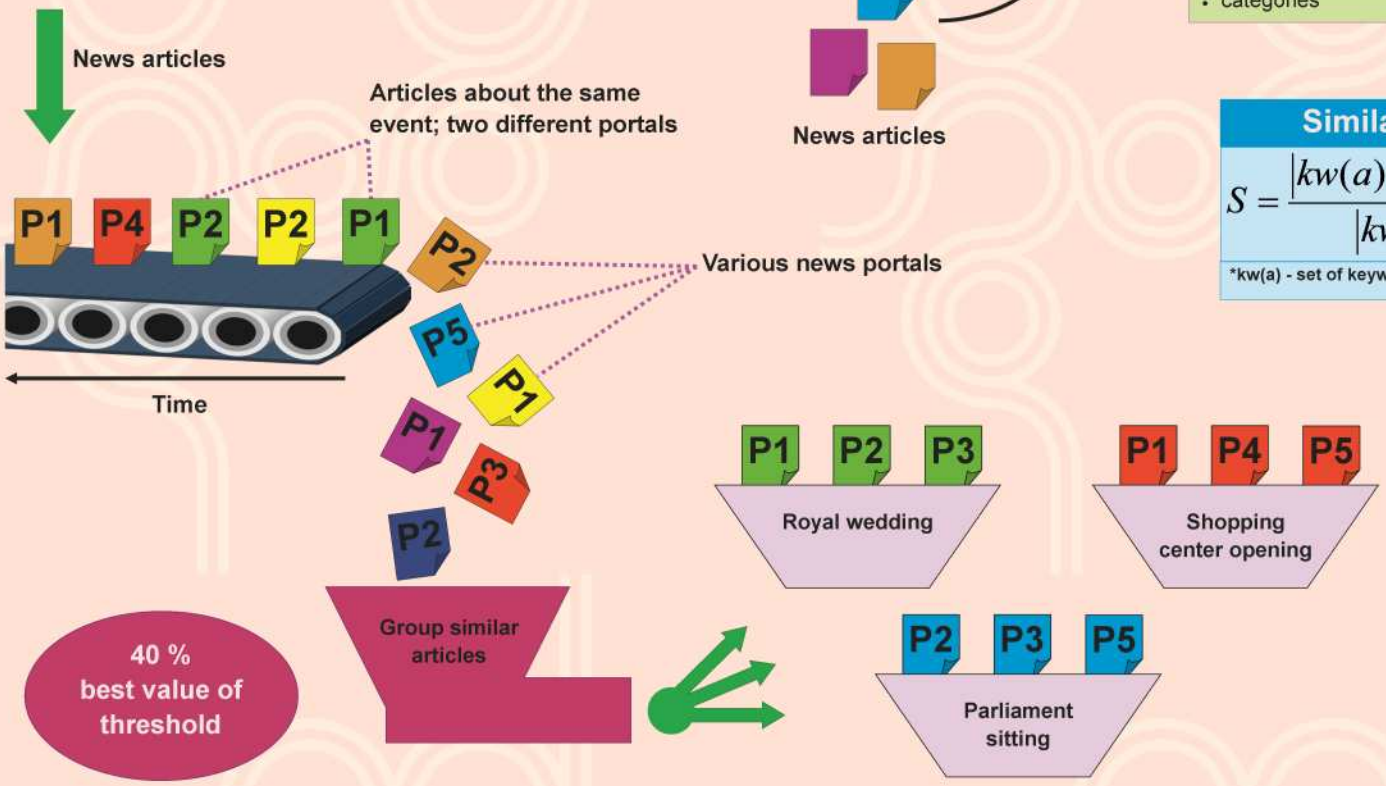
- Parliament sitting
- Royal wedding
- Shopping center opening

Articles about events published on various portals

Feature extraction

Other features

- keywords
- publication time
- term frequency
- named entities
- categories



Similarity

$$S = \frac{|kw(a) \cap kw(b)|}{|kw(a)|}$$

*kw(a) - set of keywords of article *a*

Keywords extraction

- select the main content of an article
- translate the text to English
- use KW extraction web services
  - OpenCalais.com
  - tagthe.net
  - delicious.com
- this all is done using Metall web service
  - peweproxy.fiit.stuba.sk/metall

Grouping algorithm for article *n*

```
1:  for each grouped article a do
2:      k = num of identical keywords of n and a
3:      similarity = k / number of keywords of n
4:      if similarity > max similarity then
6:          max similarity = similarity
7:          similar article = a
8:  if similar article NOT found then
9:      put n into new group
10: else put n into group of similar article
```

Results

Total groups created	46
Articles put into wrong group	4
Groups which can be further divided	1
Groups which can be joined with others	18
Articles moved to correct group	4
New correct groups made from incorrect one	12

Faculty of Informatics and Information Technologies
Slovak University of Technology, Bratislava

PeWe@FIIT
personalized web group