FiND: A Real-time Filtering by Novelty and Diversity for Publish/Subscribe Systems

SSDBM’15 - San Diego, CA, USA

Z. Hmedeh¹,², C. du Mouza², N. Travers²

(1) University of Paris Nanterre X, Nanterre, France
(2) CEDRIC Laboratory - Cnam, Paris, France
Information delivery problem

Large number of sources + high publication rates

Huge volumes of data
Problem
How users can receive only interesting information?

Issue
How can we efficiently send relevant RSS items to users?
FiND: Filtering by Novelty and Diversity

Two steps

- **Matching**: Items containing subscriptions’ terms
  - Index for broadmatch and partial match [EDBT’12]
- **Filtering**:
  - By novelty: Remove redundant information
  - By diversity: Maximize diversification in the delivered information items
- **History based**:
  - Filtering by received items by the user
  - Sliding window on time ($W = 24$ hours)
Subscription = “Football 2014”

- History
  - Football 2014 Italy
  - Football 2014 France Germany

New incoming items

<table>
<thead>
<tr>
<th></th>
<th>Novelty</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Football 2014 France</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Football 2014 France Italy</td>
<td>√</td>
<td>X</td>
</tr>
<tr>
<td>Football 2014 Italy Brazil</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Novelty and Diversity scores

Novelty scores

- Asymmetric measure ($\neq$ Jaccard)
  - Time dependent
- Text overlap
  - new terms $\rightarrow$ new information
  - Weight terms: term discrimination value (TDV)

Diversity scores

- Increase the average pairwise distance
  - New history versus the old one
  - Diversity($H \cup I - I_{old}$) > Diversity($H$)
  - Efficient computation: $O(2)$ versus $O(n)$
- Need a distance measure
  - Adapted to short items ($\neq$ Jaccard or cosine)
  $\Rightarrow$ Euclidian distance with weights
Demonstration

**Dataset**
- Real dataset of 10.7M items
- 100M Subscriptions automatically generated (Alias sampling method, with Google distribution)

**Demonstrations highlights**
- Impact of different parameters on time and result set of $S$
- Creation of $S$, set novelty threshold and diversity’s impact
- Result set’s quality: compared to the top-k algorithm
Thanks for your attention!

Questions?