SubSift: a novel application of the vector space model to support the academic research process

Simon Price
Institute for Learning and Research Technology

Peter A. Flach and Sebastian Spiegler
Intelligent Systems Laboratory
SubSift

- SubSift is a prototype application to support academic peer review.
- SubSift matches submitted conference/journal papers to potential peer reviewers based on similarity to published works.
- Website: http://subsift.ilrt.bris.ac.uk
Contribution of this work

- Innovative application of established theory
- Open Source software
- Hosted web services
- Example applications
Outline of this paper

1. Motivation and Implementation

2. Background Theory
   - Vector Space Model
   - Representational State Transfer (REST)

3. SubSift Web Services

4. Applications
1. Motivation and Implementation

2. Background Theory

3. SubSift Web Services

4. Applications
Motivation: KDD’09 review process

Peter Flach was programme committee (PC) co-chair

- 500+ papers and 200+ PC members
- Idea: streamline the paper bidding and allocation process
- Software developed to do this - part of which we named **SubSift**
  - bid initialisation (3=want to review, ..., 0=do not want)
  - papers ranked for each PC member
  - PC members ranked for each paper

**Further details:**
*Novel tools to streamline the conference review process: Experiences from SIGKDD’09*
Evaluating SubSift at KDD’09

- **Precision and recall:**
  - 88% median precision
    (non-zero actual bids among non-zero predicted bids)
  - 80% median recall
    (non-zero predicted bids among non-zero actual bids)

- **User feedback:**
  “…as I go thru my paper assignments, I am extremely impressed by quality of your initial automated assignment!”

  *Gregory Piatetsky (KDD’09 reviewer)*
Implementation

• Project to repack SubSift as web services

Submitted bid: **SubSift Services**

“Rapid Innovation” call under the JISC Information Environment Programme
Implementation

• Project to repackage SubSift as web services

Submitted bid: **SubSift Services**

“Rapid Innovation” call under the JISC Information Environment Programme
SubSift Services

Workflow Inputs

Papers

PC URIs

workflow Outputs

Workflow Outputs

PC-Papers

Paper-PCs

Bids

 工作流程 输出

工作流程 输入

PROFILE SERVICE

DOCUMENT SERVICE

Documents

Text

XML, JSON, ...

PROFILE SERVICE

DOCUMENT SERVICE

Text

HTML

BOOKMARKS SERVICE

XML

HTML

MATCH SERVICE

HARVESTER ROBOT

Text

HTML

HTML GENERATOR

HTML

HTML GENERATOR

CSV
2. Background Theory

1. Motivation and Implementation
2. Background Theory
3. SubSift Web Services
4. Applications
Vector Space Model (from Information Retrieval)

For a query \( q \), rank the documents \( d_j \) in collection \( D \) by descending similarity to the query.

Vector Space Model consists of:

- *bag-of-words* representation
- cosine similarity
- tf-idf weighting
Vector Space Model: *bag-of-words* representation

### no. terms in each abstract

<table>
<thead>
<tr>
<th></th>
<th>intelligence</th>
<th>learning</th>
<th>logic</th>
<th>machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstract 1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>abstract 2</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>abstract 3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

### no. terms in DBLP author page of each PC member

<table>
<thead>
<tr>
<th></th>
<th>intelligence</th>
<th>learning</th>
<th>logic</th>
<th>machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc member 1</td>
<td>10</td>
<td>70</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>pc member 2</td>
<td>0</td>
<td>70</td>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>pc member 3</td>
<td>30</td>
<td>70</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Vector Space Model: cosine similarity

Query and document similar if angle $\theta$ between their vectors is small.

$$\text{similarity}_{\text{cosine}}(\vec{q}, \vec{d}) = \cos(\theta) = \frac{\vec{q} \cdot \vec{d}}{||\vec{q}|| \cdot ||\vec{d}||}$$

- $d \in D$ – document represented as multiset of terms (bag-of-words).
- $\vec{d}$ – document vector in the vector space defined by vocabulary of $D$.
- $\vec{q}$ – query vector in the same vector space as $\vec{d}$.
Vector Space Model: tf-idf weighting

Normalise term counts within document and penalise common terms in $D$.

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad idf_i = \log_2 \left( \frac{|D|}{df_i} \right), \quad tf-idf_{ij} = tf_{ij} \times idf_j$$

- $tf_{ij}$ is term frequency of term $t_i$ in the document $d_j$.
- $n_{ij}$ is term count, the number of times term $t_i$ occurs in the document $d_j$.
- $df_i$ is document frequency of term $t_i$ is the number of documents in $D$ in which term $t_i$ occurs.
Vector Space Model: \( \text{tf-idf weighting} \)

Normalise term counts within document and penalise common terms in \( D \).

\[
\text{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad \text{idf}_i = \log_2 \left( \frac{|D|}{\text{df}_i} \right), \quad \text{tf-idf}_{ij} = \text{tf}_{ij} \times \text{idf}_j
\]

- \( \text{tf}_{ij} \) is term frequency of term \( t_i \) in the document \( d_j \).
- \( n_{ij} \) is term count, the number of times term \( t_i \) occurs in the document \( d_j \).
- \( \text{df}_i \) is document frequency of term \( t_i \) is the number of documents in \( D \) in which term \( t_i \) occurs.

<table>
<thead>
<tr>
<th>intelligence</th>
<th>learning</th>
<th>logic</th>
<th>machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc member 1</td>
<td>10</td>
<td>70</td>
<td>20</td>
</tr>
<tr>
<td>pc member 2</td>
<td>0</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>pc member 3</td>
<td>30</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>
Representational State Transfer (REST)

REST is a design pattern for **web services** based on HTTP using its familiar URIs, requests, responses, authentication, etc.

“RESTful” web services:

- URIs to represent resources
- HTTP POST/GET/PUT/DELETE correspond to usual Create/Read/Update/Delete (CRUD) operations
- Response formats typically include: XML, JSON, CSV
SubSift Web Services

1. Motivation and Implementation
2. Background Theory
3. SubSift Web Services
4. Applications
3. Matches

In SubSift, a match item is a similarity score (and supporting statistics) representing how alike a specific pair of profile items are. Each matches folder is a container to hold a list of match items. A matches folder is created by analysing every pairing of profile items drawn from a pair of profiles folders. Each match item scores the similarity of a single profile from the first profiles folder against every profile from the second profiles folder. This allows the users to form clusters of people and to search for similar people both within and between different profiles. There is a common use case for a large conference to produce a list of potential reviewers (one per paper) based on a comparison with the bibliography pages of programme committee members in order to rank potential reviewers for each paper and vica versa.

### 3.1 MATCHES FOLDERS

<table>
<thead>
<tr>
<th>API Method</th>
<th>HTTP Method</th>
<th>URI Schema</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>matches list</td>
<td>GET</td>
<td>/user/matches</td>
<td>description, mode, limit, threshold, sort, full</td>
</tr>
<tr>
<td>matches show</td>
<td>GET</td>
<td>/user_id/matches/sterol_id</td>
<td>description, mode, limit, threshold, sort, full</td>
</tr>
<tr>
<td>matches exist</td>
<td>POST</td>
<td>/user_id/matches/sterol_id/profile_id/profile_id</td>
<td>description, mode, limit, threshold, sort, full</td>
</tr>
<tr>
<td>matches create</td>
<td>POST</td>
<td>/user_id/matches/sterol_id/profile_id/profile_id</td>
<td>description, mode, limit, threshold, sort, full</td>
</tr>
<tr>
<td>matches update</td>
<td>PUT</td>
<td>/user_id/matches/sterol_id/profile_id/profile_id</td>
<td>description, mode, limit, threshold, sort, full</td>
</tr>
</tbody>
</table>
Applications

1. Motivation and Implementation
2. Background Theory
3. SubSift Web Services
4. Applications
SubSift has been used for...
Profiling a research group by its publications

Diagram produced in Wordle using SubSift profile data
Finding an expert

ILRT Matcher

Enter a title, abstract or text of a paper and click Submit to compare against a pre-defined set of profiles.

Text:
Semantic Web technologies have moved beyond the point of being promising futuristic technologies and demonstration projects, to being technologies in action in realistic contexts and conditions. Semantic Web applications are being developed for many aspects of scientific research, from experimental data management, discovery and retrieval, to analytic workflows, hypothesis development and testing, to research publishing and dissemination. This workshop intends to explore the questions that arise as Semantic Web applications are increasingly grounded within the actual lifecycle of scientific research, from observation and hypothesis formulation to publication, dissemination and criticism. We aim to bring together researchers across the disciplines, to discuss the use, development and embedding of these technologies in varied research domains and contexts. We will discuss the actuality of Semantic Web technologies in use and the emergent practices through which they are being developed and deployed. We aim to encourage vigorous discussion around aims, methods, applications and pragmatics. This workshop will look at the theoretical, methodological and pragmatic issues of grounding the development, deployment and evolution of ontologies and applications in Semantic e-

Submit

ILRT staff ranked by similarity to text

<table>
<thead>
<tr>
<th>Name</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Damian Steer</td>
<td>0.142</td>
</tr>
<tr>
<td>Mike Jones</td>
<td>0.124</td>
</tr>
<tr>
<td>Nikki Rogers</td>
<td>0.095</td>
</tr>
<tr>
<td>Jasper Tredgold</td>
<td>0.072</td>
</tr>
<tr>
<td>Sarah Agarwal</td>
<td>0.059</td>
</tr>
<tr>
<td>Simon Price</td>
<td>0.057</td>
</tr>
<tr>
<td>Ben Joyner</td>
<td>0.045</td>
</tr>
<tr>
<td>Paul Shabjdee</td>
<td>0.045</td>
</tr>
<tr>
<td>Chris Bailey</td>
<td>0.043</td>
</tr>
</tbody>
</table>
Clustering staff based on homepage similarity

Dendrogram produced in Matlab from SubSift generated similarity matrix
Precision-recall at different thresholds
Similarity networks

Diagram created by Graphvis from SubSift generated dot file
Connectivity

Diagram created by Graphvis from SubSift generated dot file
And finally...
Conclusion

Repackaging SubSift as SubSift Services

- Created a more general purpose resource
- Potential applications outside of peer review domain

Publishing functionality as web services

- Similar approach may work for other research-produced applications