



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

**Simon Price**

Institute for Learning and Research Technology



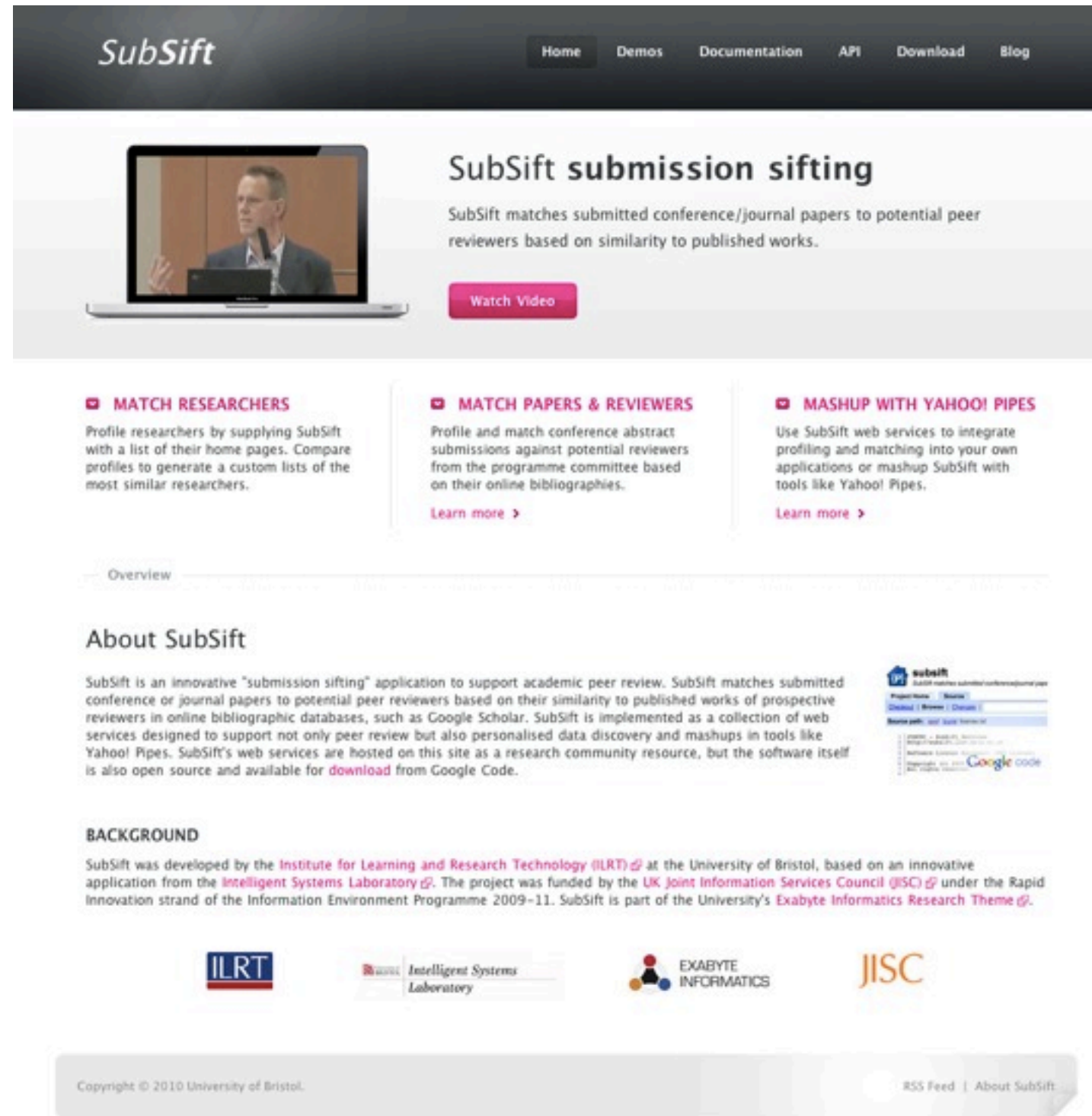
**Peter A. Flach and Sebastian Spiegler**

Intelligent Systems Laboratory







*Intelligent Systems  
Laboratory*

- 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>



The screenshot shows the SubSift website interface. At the top, there is a navigation bar with links for Home, Demos, Documentation, API, Download, and Blog. The main heading is "SubSift submission sifting", followed by a sub-heading "SubSift matches submitted conference/journal papers to potential peer reviewers based on similarity to published works." Below this is a "Watch Video" button. Three feature boxes are displayed: "MATCH RESEARCHERS" (profile researchers by supplying SubSift with a list of their home pages), "MATCH PAPERS & REVIEWERS" (profile and match conference abstract submissions against potential reviewers), and "MASHUP WITH YAHOO! PIPES" (use SubSift web services to integrate profiling and matching into your own applications). The page also includes an "About SubSift" section, a "BACKGROUND" section, and logos for ILRT, Intelligent Systems Laboratory, EXABYTE INFORMATICS, and JISC. The footer contains copyright information for the University of Bristol and links for RSS Feed and About SubSift.

# 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



# I. Motivation and Implementation

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:

Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver Ralf, Herbrich, Thore Graepel, and Mohammed J. Zaki.

*Novel tools to streamline the conference review process: Experiences from SIGKDD'09*  
SIGKDD Explorations, 11(2):63–67, December 2009

# 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 repackage SubSift as web services



Submitted bid: **SubSift Services**



“Rapid Innovation” call under the JISC Information Environment Programme



# Implementation

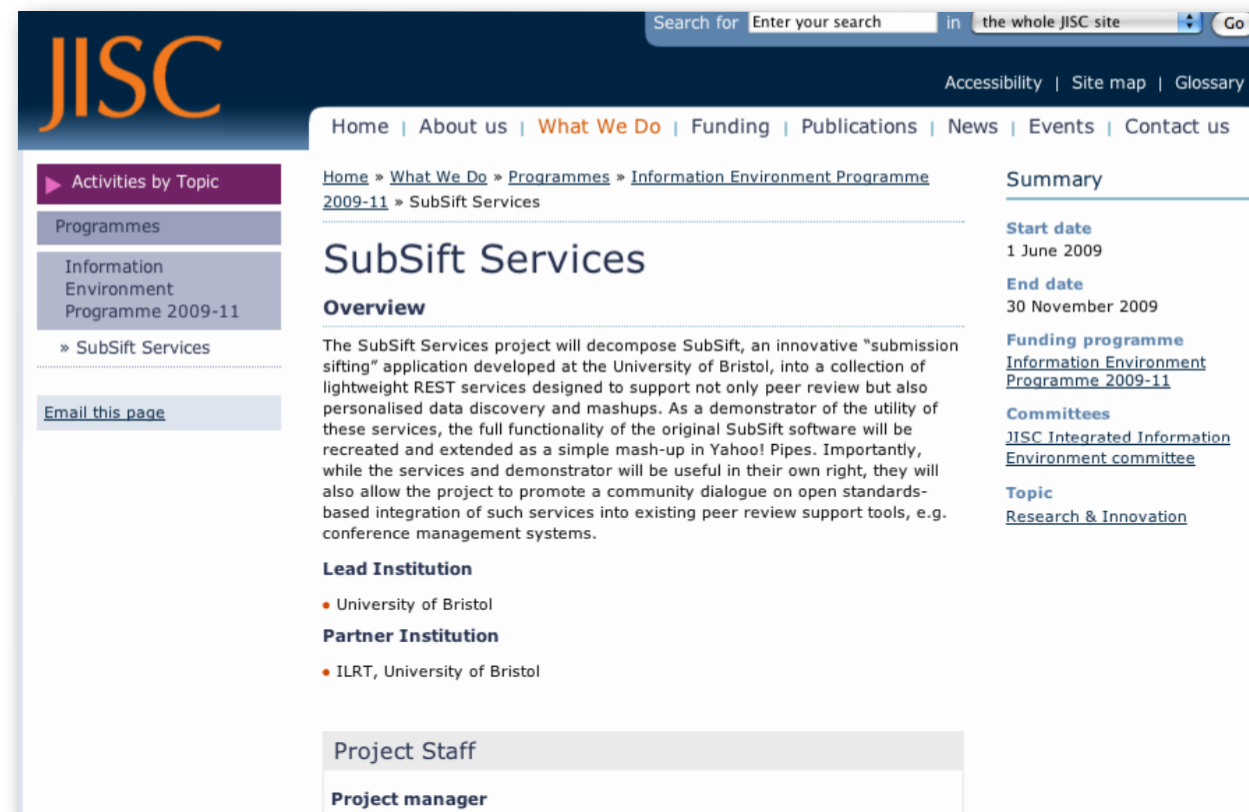
- Project to repackage SubSift as web services



Submitted bid: **SubSift Services**



“Rapid Innovation” call under the JISC Information Environment Programme



The screenshot shows the JISC website interface. At the top, there is a search bar and navigation links for Accessibility, Site map, and Glossary. The main content area is titled "SubSift Services" and includes an overview, lead institution (University of Bristol), partner institution (ILRT, University of Bristol), and project staff information. A summary sidebar on the right provides key dates and program details.

**Summary**

- Start date:** 1 June 2009
- End date:** 30 November 2009
- Funding programme:** Information Environment Programme 2009-11
- Committees:** JISC Integrated Information Environment committee
- Topic:** Research & Innovation

**Lead Institution**

- University of Bristol

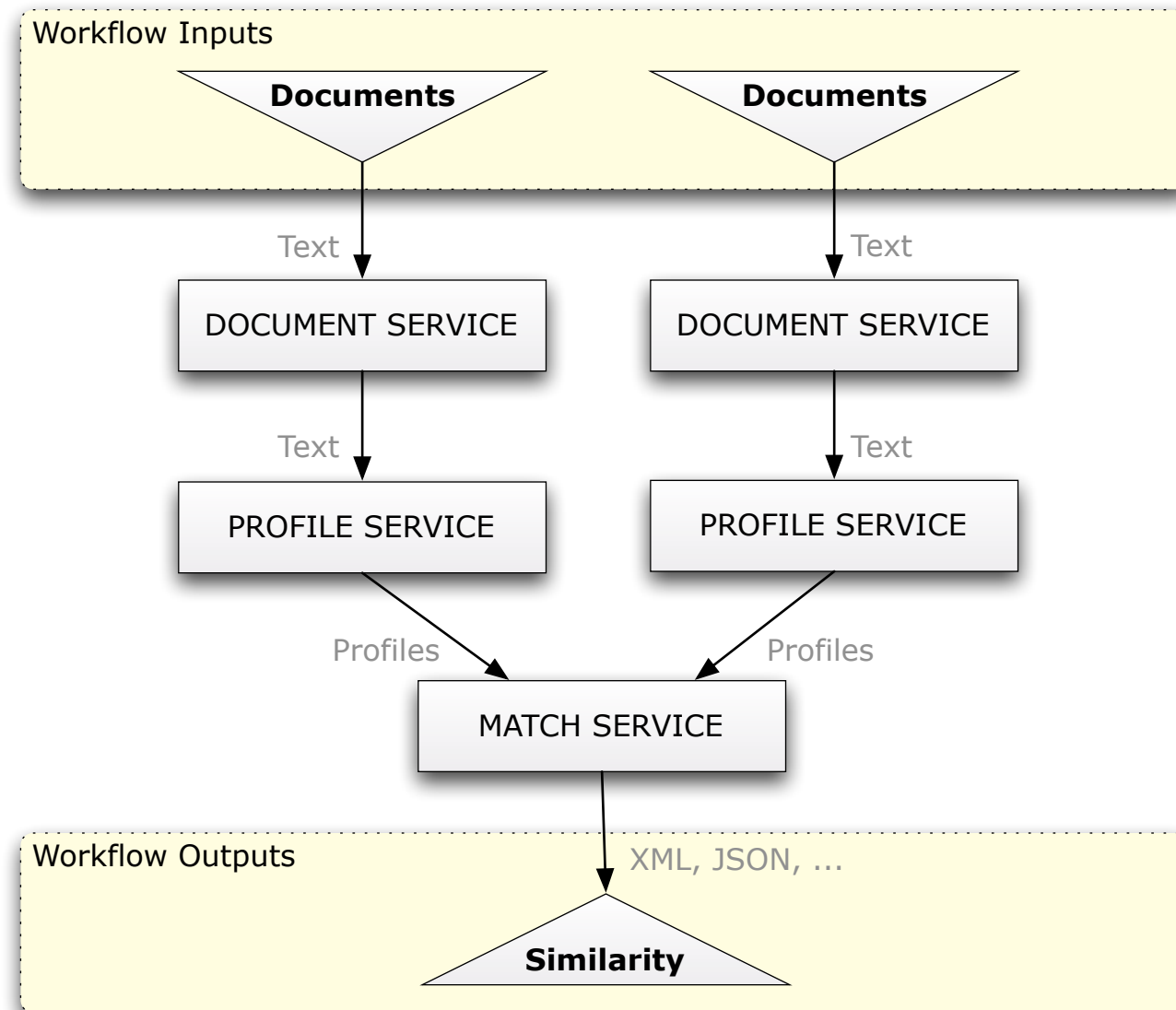
**Partner Institution**

- ILRT, University of Bristol

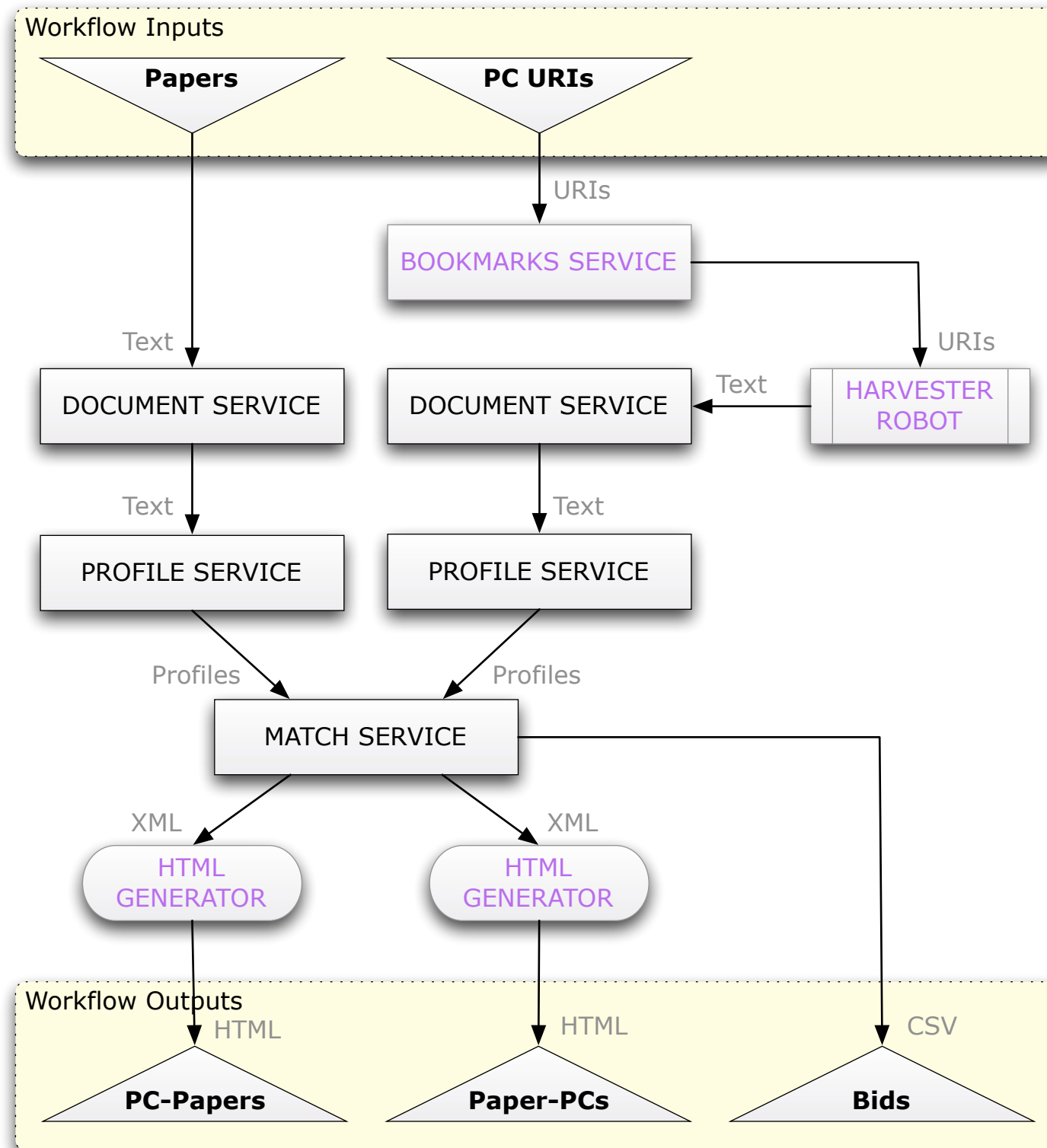
**Project Staff**

**Project manager**

# SubSift Services



# SubSift Services



## 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

	<b>intelligence</b>	<b>learning</b>	<b>logic</b>	<b>machine</b>
<b>abstract 1</b>	0	2	0	1
<b>abstract 2</b>	3	0	5	0
<b>abstract 3</b>	2	1	0	2

no. terms in DBLP author page of each PC member

	<b>intelligence</b>	<b>learning</b>	<b>logic</b>	<b>machine</b>
<b>pc member 1</b>	10	70	20	0
<b>pc member 2</b>	0	70	5	99
<b>pc member 3</b>	30	70	0	0



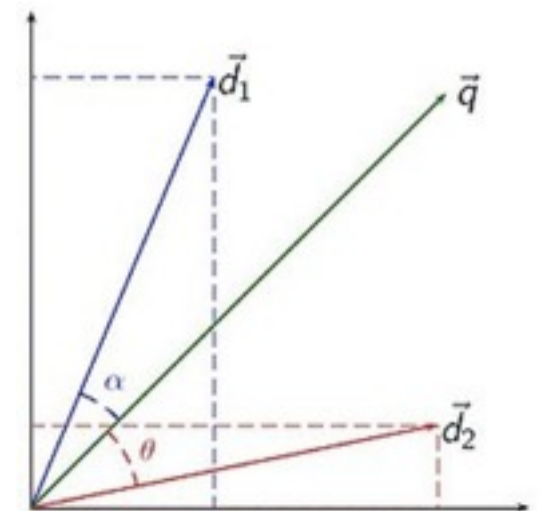


# 🔥 Vector Space Model: cosine similarity

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  $D$ .

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

- $\text{tf}_{ij}$  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.





# Vector Space Model: 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.

	intelligence	learning	logic	machine
pc member 1	10	70	20	0
pc member 2	0	70	5	99
pc member 3	30	70	0	0

# 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

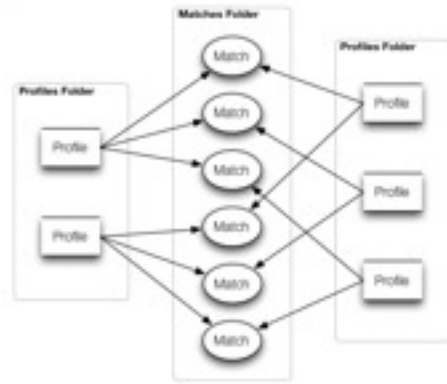
# SubSift Web Services

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

# SubSift REST API

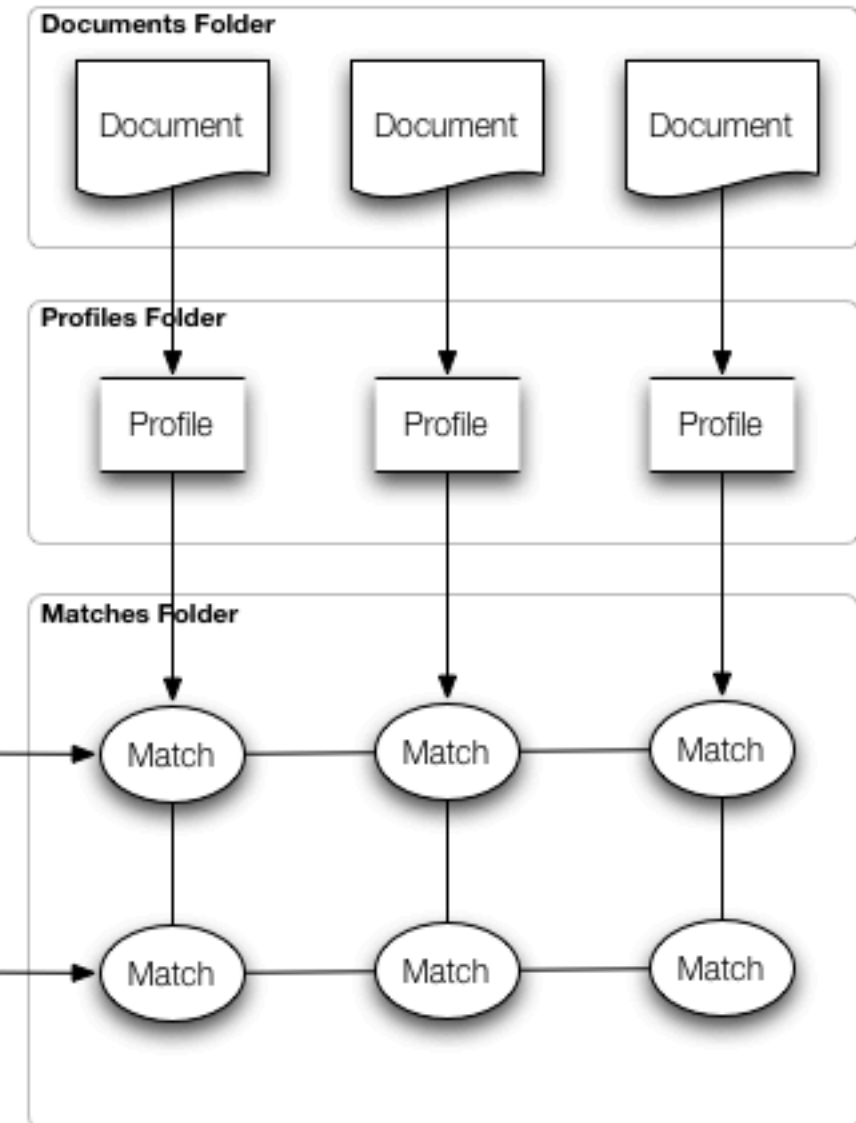
## 3. Matches

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. A typical usage of such a comparison is to match submitted conference abstracts with the bibliography pages of programme committee members in order to rank potential reviewers for each paper and visa versa.



### 3.1 MATCHES FOLDERS

API Method	HTTP	URI Schema	Parameters
matches list	GET	/confer_id/matches	sort, full
matches show	GET	/confer_id/matches/cfolder_id	sort, full
matches exists	HEAD	/confer_id/matches/cfolder_id	
matches create	POST	/confer_id/matches/cfolder_id/profiles/profile_id1/+with/profile_id2	description, mode, limit, threshold, sort, full
	POST	/confer_id/matches/cfolder_id	profile_id1, profile_id2, description, mode, limit, threshold, sort, full
matches update	PUT	/confer_id/matches/cfolder_id/profiles/profile_id1/+with/profile_id2	description, mode, limit, threshold, sort, full





# Applications

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

# SubSift has been used for..

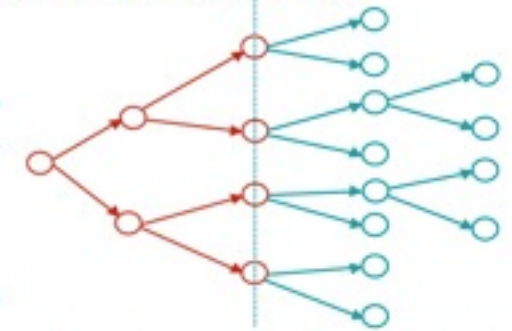


ICDM 2010: The 10th IEEE International Conference on Data Mining  
December 13-17, 2010, Sydney, Australia



2010 SIAM International Conference on **DATA MINING**

April 29-May 1, 2010



**Profiling with SubSift and Subgroup Discovery**  
Cheng Peng, Srikumar S. Iyengar, Peter Floukas & Steve Price  
Department of Computer Science, University of Bristol

**Objectives**  
This paper presents a novel approach to subgroup discovery. It introduces a new algorithm, SubSift, which is designed to find subgroups that are highly discriminative for a given target class. The algorithm is based on a novel notion of subgroup quality, which is defined in terms of the ratio of the number of instances in the subgroup that belong to the target class to the total number of instances in the subgroup. The algorithm is designed to find subgroups that are highly discriminative for a given target class, and it is able to handle both binary and multi-class classification problems.

**Technical Overview**  
The algorithm is based on a novel notion of subgroup quality, which is defined in terms of the ratio of the number of instances in the subgroup that belong to the target class to the total number of instances in the subgroup. The algorithm is designed to find subgroups that are highly discriminative for a given target class, and it is able to handle both binary and multi-class classification problems.

**Experimental Results**

Dataset	Method	Quality	Runtime
Adult	SubSift	0.85	10s
	Other	0.75	5s
Census	SubSift	0.90	15s
	Other	0.80	10s
Census-Adult	SubSift	0.88	20s
	Other	0.78	15s
Census-Adult-Sub	SubSift	0.92	25s
	Other	0.82	20s

**Conclusions**  
This paper presents a novel approach to subgroup discovery. It introduces a new algorithm, SubSift, which is designed to find subgroups that are highly discriminative for a given target class. The algorithm is based on a novel notion of subgroup quality, which is defined in terms of the ratio of the number of instances in the subgroup that belong to the target class to the total number of instances in the subgroup. The algorithm is designed to find subgroups that are highly discriminative for a given target class, and it is able to handle both binary and multi-class classification problems.





# 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

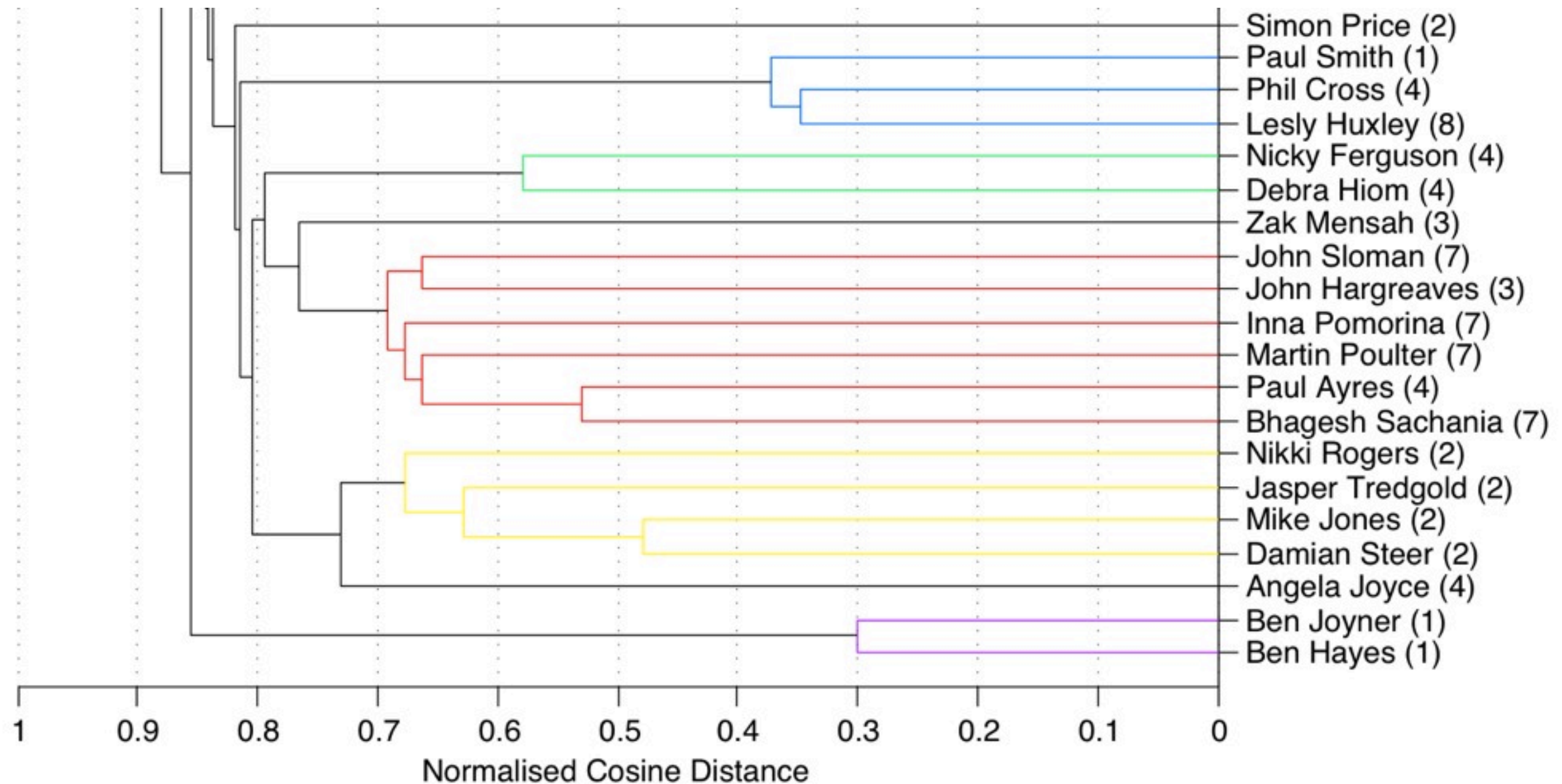
### ILRT staff ranked by similarity to text

Damian Steer	0.142																			
Mike Jones	0.124																			
Nikki Rogers	0.095																			
Jasper Tredgold	0.072																			
Sarah Agarwal	0.059																			
Simon Price	0.057																			
Ben Joyner	0.045																			
Paul Shabajee	0.045																			
Chris Bailey	0.043																			



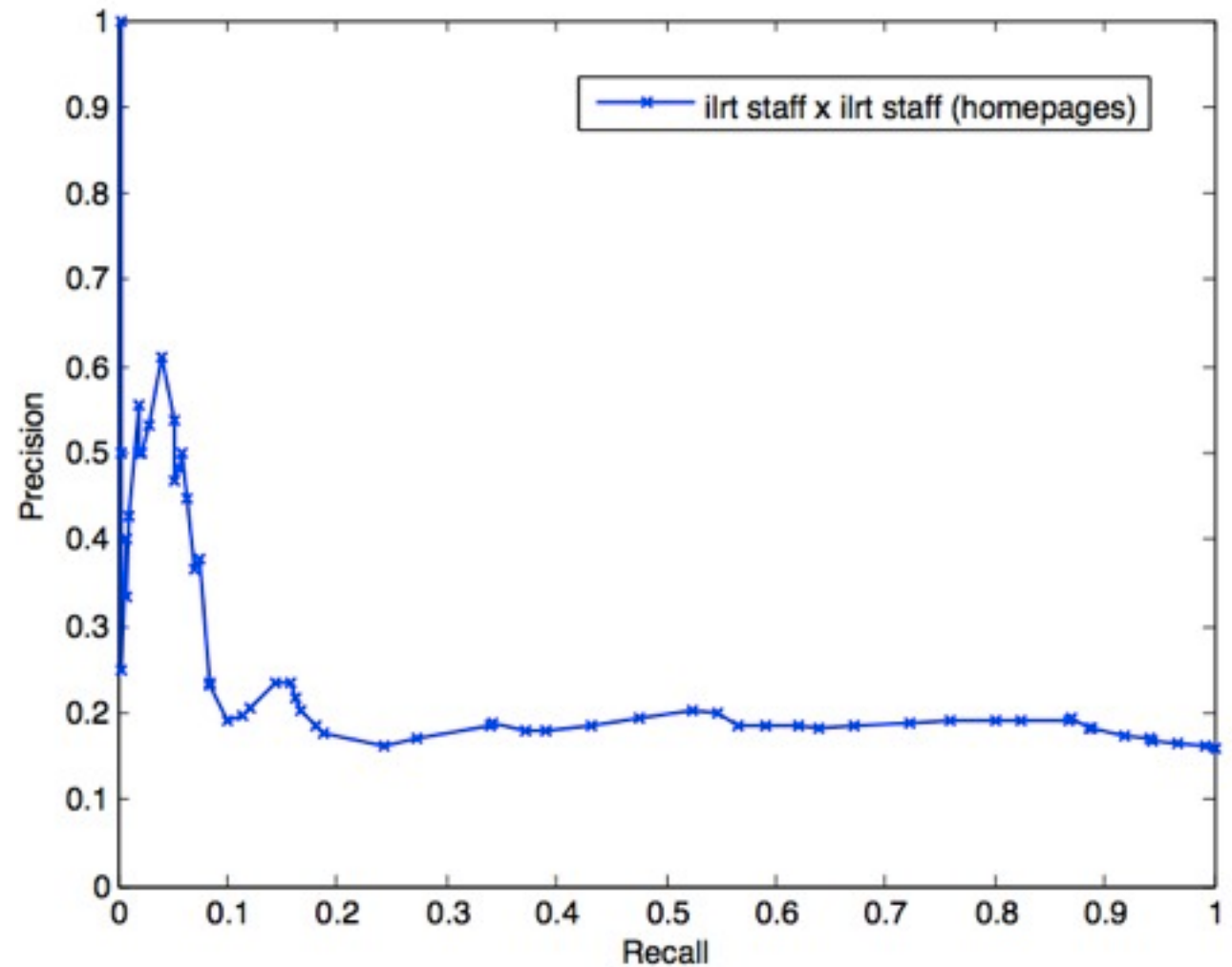
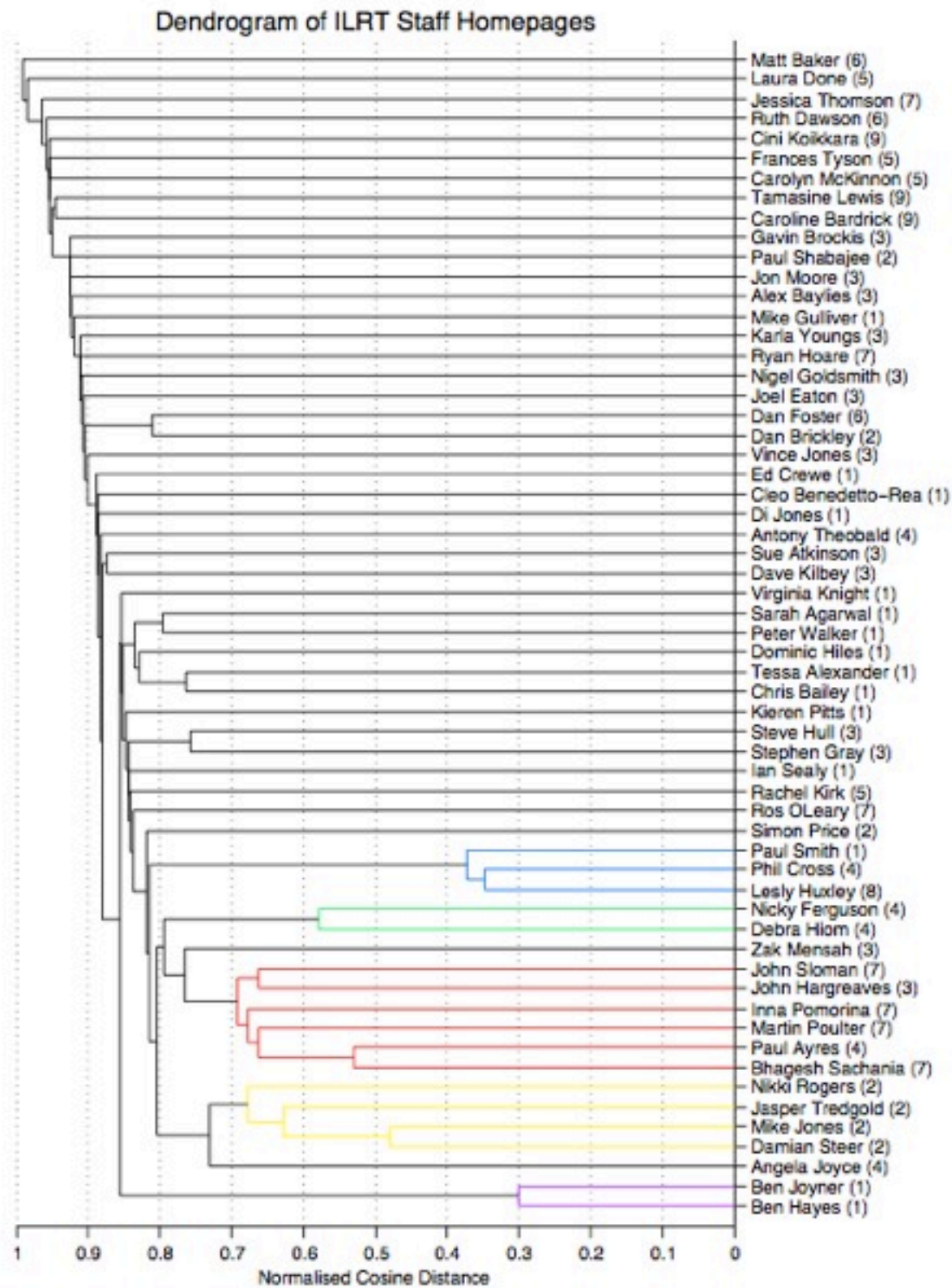


# 🔥 Clustering staff based on homepage similarity



Dendrogram produced in Matlab from SubSift generated similarity matrix

# Precision-recall at different thresholds







# Connectivity

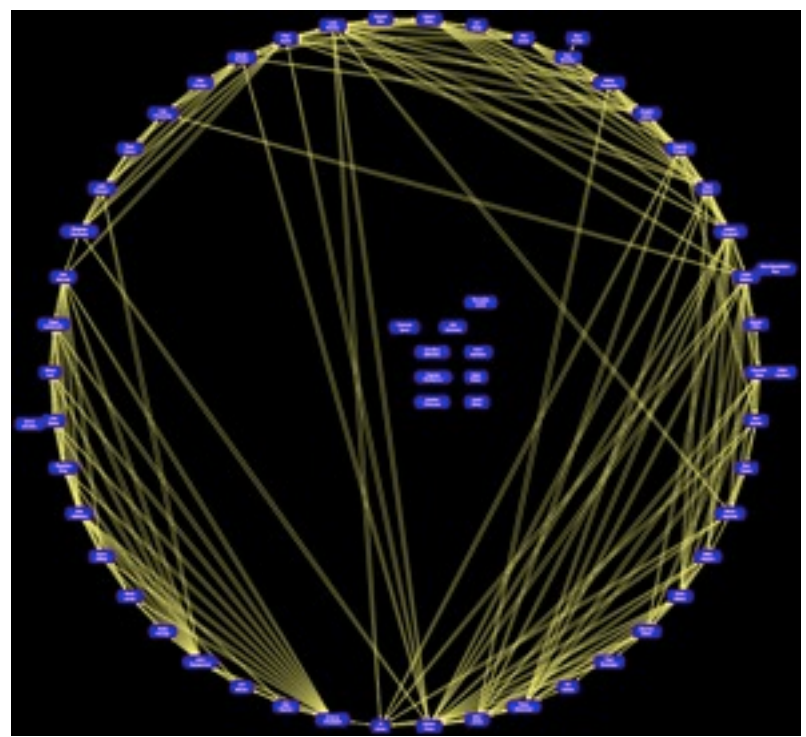
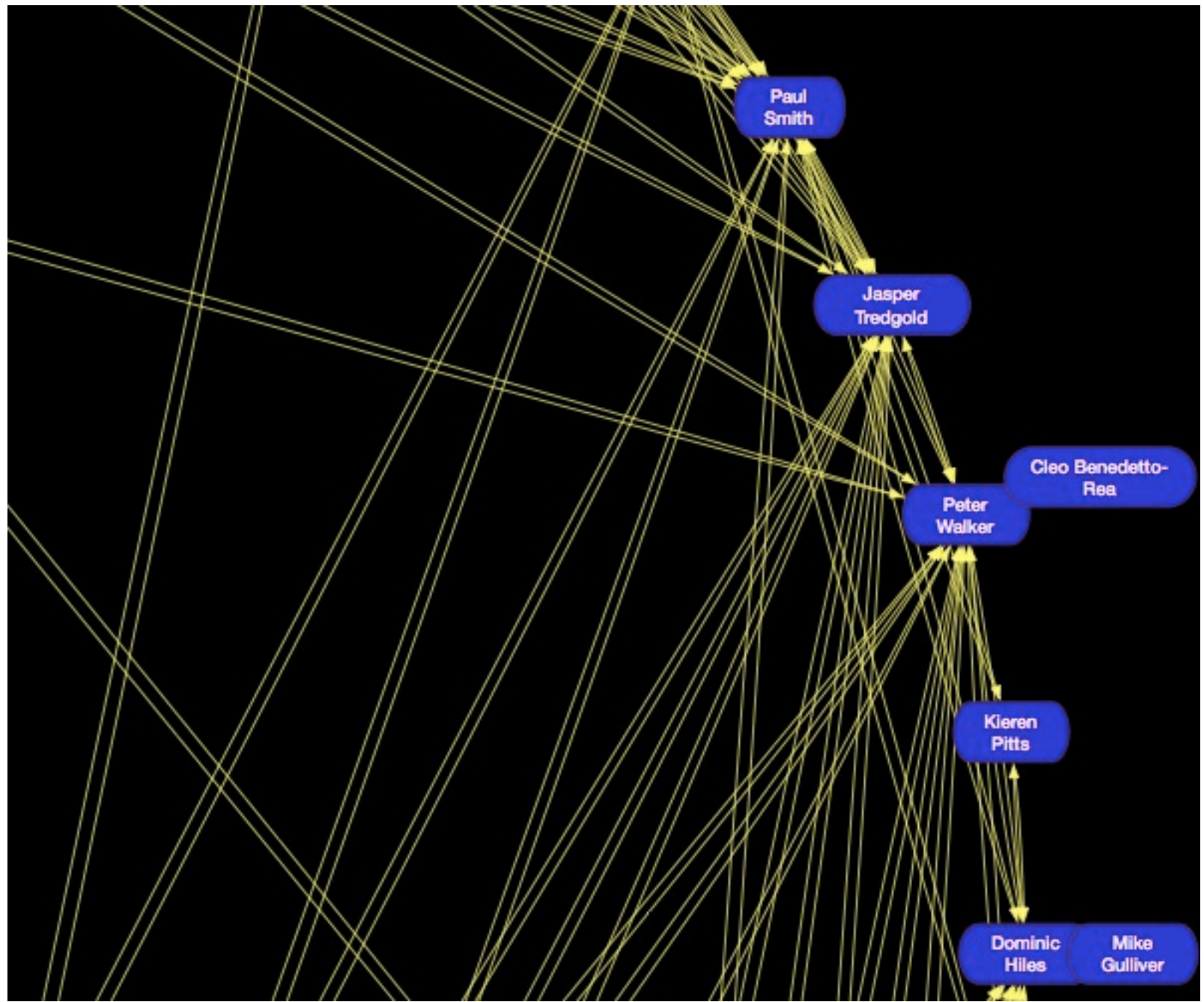


Diagram created by Graphvis from SubSift generated dot file

 **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

