

# Iterative Entity Navigation via Co-clustering Semantic Links and Entity Classes

Liang Zheng, Jiang Xu, Jidong Jiang, Yuzhong Qu, Gong Cheng  
Nanjing University, China



# Navigation is everywhere

Physical world



Web-browsing



# Using hyperlinks for navigation on the Web

## Greece

From Wikipedia, the free encyclopedia

Coordinates:  39° N 22° E

For other uses, see *Greece (disambiguation)*.

**Greece** (Greek: Ελλάδα, *Elláda* [ˈlaða]  (ⓘ) (ⓘ) listen)), officially the **Hellenic Republic** (Greek: Ελληνική Δημοκρατία *Ellīnikī́ Dīmokratía* [eliniˈci ðimokraˈti.a]  (ⓘ) (ⓘ)), also known since ancient times as **Hellas** (Ancient Greek: Ἑλλάς *Ellás*



## Acropolis of Athens

From Wikipedia, the free encyclopedia

Coordinates:  37.971421° N 23.726166° E

For the neighbourhood of Athens, see *Acropolis (neighbourhood)*.

The **Acropolis of Athens**

(Ancient Greek:

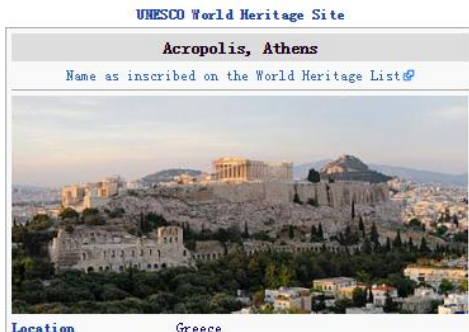
Ἀκρόπολις;<sup>[1]</sup>

Modern Greek:

Ἀκρόπολη

Ἀθῆναι *Akropoli*

*Athinai*) is an ancient citadel located on a extremely rocky outcrop above the city of Athens and contains the remains of several ancient buildings of great



## Athens

From Wikipedia, the free encyclopedia

Coordinates:  37° 58′ N 23° 43′ E

*This article is about the capital city of Greece. For other uses, see *Athens (disambiguation)*.*

**Athens** (/ˈæθɪnz/ [2] Modern Greek:

Αθήνα, *Athína* Greek

pronunciation: [aˈθina],

Ancient Greek: Ἀθῆναι *Athēnai*), is the capital and largest city of Greece.

Athens dominates the Attica region and is one of the world's oldest cities, with its recorded history spanning around 3,400 years, and the earliest human presence started somewhere between the 11th and 7th millennia BC.<sup>[3]</sup> Classical Athens was a powerful



# Faceted navigation

## Flamenco Fine Arts Search

Images from the Collections of the Fine Arts Museums of San Francisco;  
Legion of Honor and de Young Museums, <http://www.thinker.org>

Powered by Flamenco

[Save Search](#) [History and Settings](#) [Return to Search](#) [New Search](#) [Logout](#)

☒ all items ☐ in current results

Refine your search within these categories:

**MEDIA:** [all](#) > [Drawing](#) > chalk

**LOCATION** [\(group results\)](#)

[Europe](#) (27) [North America](#) (1)

**OBJECTS:** [all](#) > [Clothing](#) > headgear

[bonnet](#) (3) [turban](#) (4)  
[hat](#) (20) [wig](#) (3)  
[headdress](#) (2)

**BUILT PLACES** [\(group results\)](#)

[Building](#) (2) [Part of Building](#) (3)  
[Built Open Space](#) (1) [Road](#) (4)  
[Dwelling](#) (1)

**ANIMALS AND PLANTS** [\(group results\)](#)

[Fish and Molluscs](#) (1) [Mammals, Other](#) (6)  
[Flowers](#) (1) [Parts of Plants](#) (1)  
[Insects](#) (3) [Trees](#) (2)  
[Mammals, Hoofed](#) (6)

**HEAVEN AND EARTH** [\(group results\)](#)

[Mountains, Hills, Valleys](#) (2) [Storms, Clouds, Floods](#) (1)  
[Rivers, Lakes, Seas](#) (6) [Sun, Moon, Stars](#) (2)

**SHAPES AND COLORS** [\(group results\)](#)

[Color](#) (7) [Scene](#) (12)  
[Decoration](#) (5) [Shape](#) (1)

**OCCUPATIONS** [\(group results\)](#)

[Combatant, Guard](#) (2) [Leader](#) (9)  
[Entertainer](#) (7) [Worker](#) (7)


These terms define your current search. Click the  to remove a term.

**MEDIA:** [Drawing](#) > chalk


**OBJECTS:** [Clothing](#) > headgear

28 items, grouped by OBJECTS ([view ungrouped items](#))


**bonnet** (3)



[Visit to Grandfather](#)  
18th century




[Study of Seated Man...](#)  
18th century




[Woman in a Bonnet](#)  
19th - 20th century


**hat** (20)




[A Chanté dans les ...](#)  
mid 19th century



[Shepherds](#)  
16th century



[Recto: Landscape; V...](#)  
circa 1870 - 1880



[Diana and Endymion](#)  
18th century



# Entity Navigation

- Entity navigation over RDF data can help users find the related entities from current browsing entities.

**OPENLINK SOFTWARE**

**About: Steven Spielberg** [Goto Sponge](#) [NotDist](#)  
An Entity of Type : [yago:SpecialEffectsPeople](#), within Data Space  
Type: [yago:SpecialEffectsPeople](#)

Attributes	Values
is <b>producer</b> of	<a href="#">The Hundred-Foot Journey</a> <a href="#">The Land Before Time</a> <a href="#">Men in Black</a> <a href="#">Poltergeist (Film)</a> <a href="#">Into the West (miniseries)</a> <a href="#">»more»</a>
is <b>relatives</b> of	<a href="#">Jessica Capshaw</a>
is <b>spouse</b> of	<a href="#">Amy Irving</a>
is <b>starring</b> of	<a href="#">Your Studio and You</a> <a href="#">Chambre 666</a> <a href="#">The Cutting Edge: The Magic of Movie Editing</a> <a href="#">Directed by John Ford</a>
is <b>story</b> of	<a href="#">The Goonies</a> <a href="#">Ace Eli and Rodger of the Skies</a>
is <b>writer</b> of	<a href="#">The Dig</a> <a href="#">Amblin'</a> <a href="#">Medal of Honor: Allied Assault</a>

**aemoo**

**Immanuel Kant**  
Philosopher  
[http://en.wikipedia.org/wiki/Immanuel\\_Kant](http://en.wikipedia.org/wiki/Immanuel_Kant)  
Café

Immanuel Kant (22 April 1724 – 12 February 1804) was a German philosopher from Königsberg, researching, lecturing and writing on philosophy and anthropology during and at the end of the 18th Century (Enlightenment). At the time, there were major successes and advances in physical science using reason and logic. But this stood in sharp contrast to the scepticism and lack of agreement or progress in empiricist philosophy. Kant's magnum opus, the Critique of Pure Reason, aimed to unite reason with experience to move beyond what he took to be failures of traditional philosophy and metaphysics. He hoped to end an age of speculation where objects outside experience were used to support what he saw as futile theories, while opposing the scepticism and idealism of thinkers such as Descartes, Berkeley and Hume. He said that 'is always remains a scandal of philosophy and universal human reason that the existence of things outside us ... should have to be assumed merely on faith, and that if it ... (continue)

David Hilbert → Königsberg → Immanuel Kant

DBpedia relations between Philosopher and City (with their frequencies):  
<http://dbpedia.org/property/birthPlace>: 66.67 %  
<http://dbpedia.org/property/placeOfBirth>: 33.33 %

Aemoo (Alberto Musetti et al 2011)  
Grouping related entities by type

<http://dbpedia.org/fct/>

## Challenges:

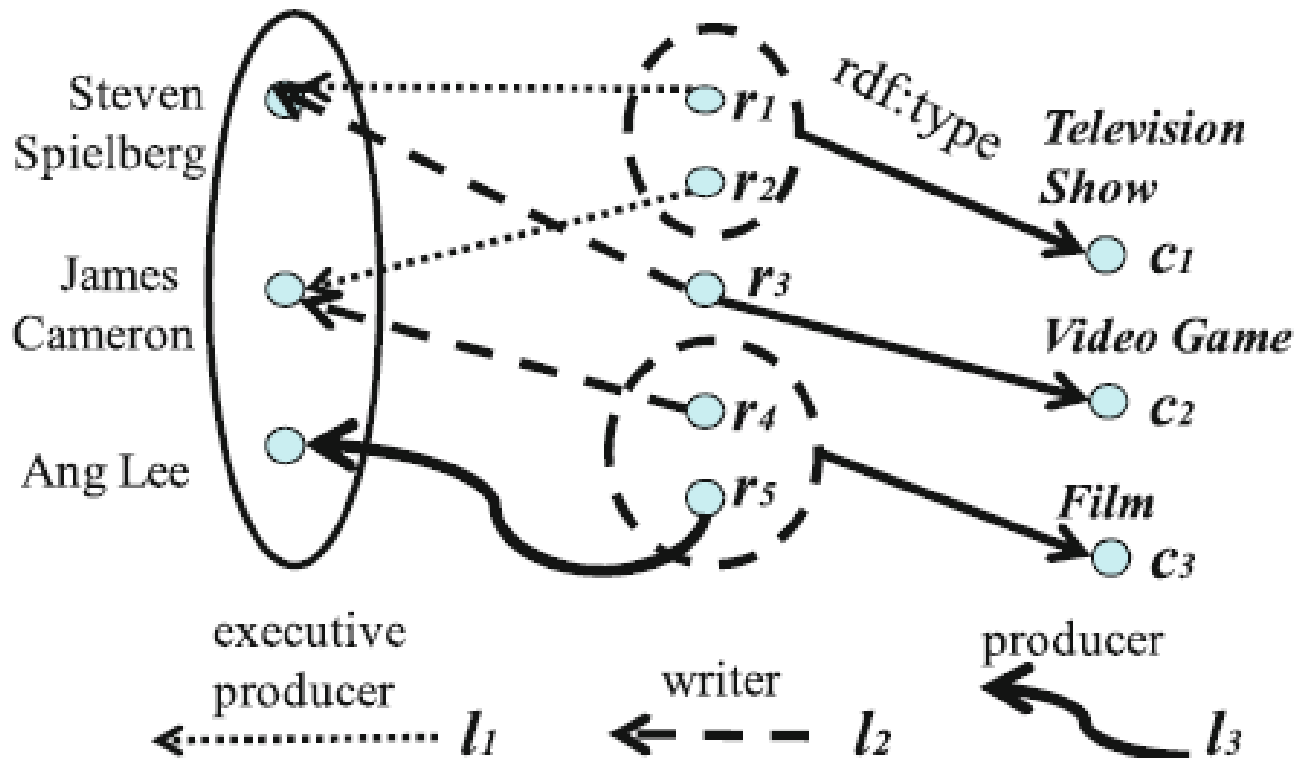
- **Large numbers** of linked entities and **high diversity** of links among entities, often make it hard for users to explore and find the entities of interest quickly.

## Problem :

- How to improve the efficiency of entity navigation?

# Motivation

- As semantic links and classes of linked entities are two key aspects to help users navigation, clustering links and classes can offer effective ways to navigate over RDF data.

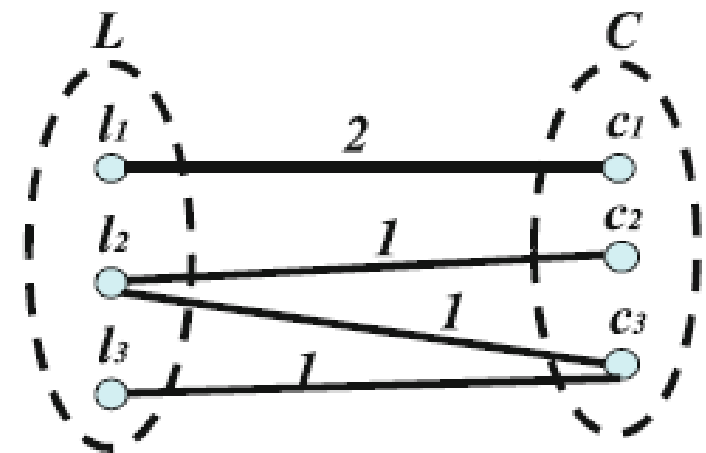
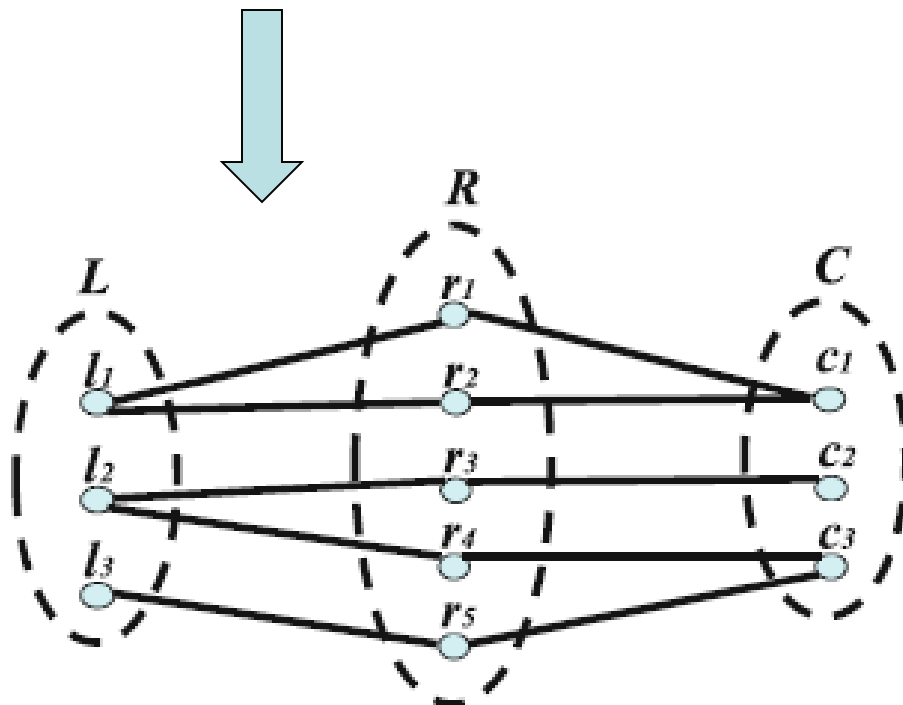
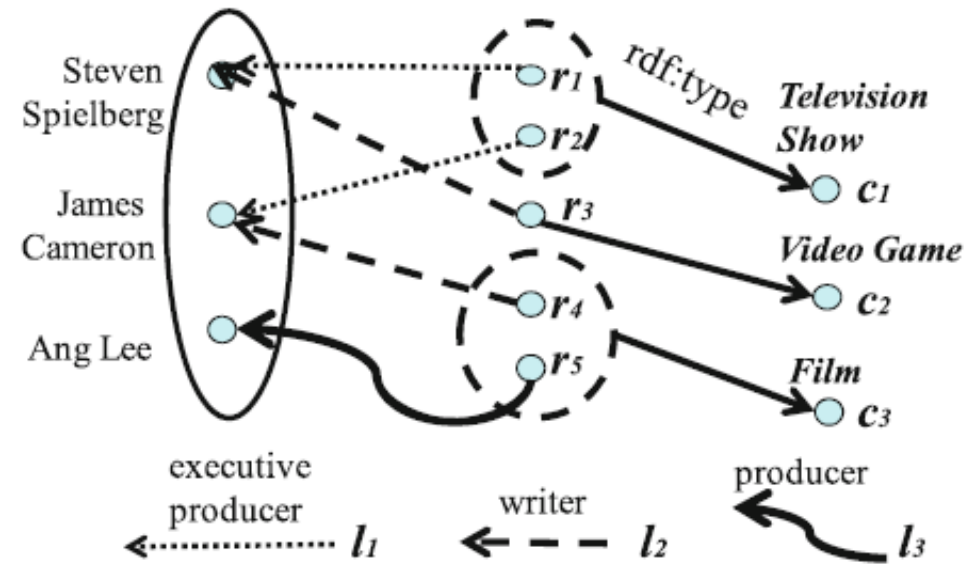


# Approach

- We propose a co-clustering approach to provide users with iterative entity navigation.
- It clusters both links and classes simultaneously utilizing both the relationship between link and class, and the intra-link relationship and intra-class relationship.



# Step 1: Build the association between links and classes



link-class graph

tripartite graph

## Step 2: Information-Theoretic co-clustering links and classes

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

# Step 2: Information-Theoretic co-clustering links and classes

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

Y

X

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

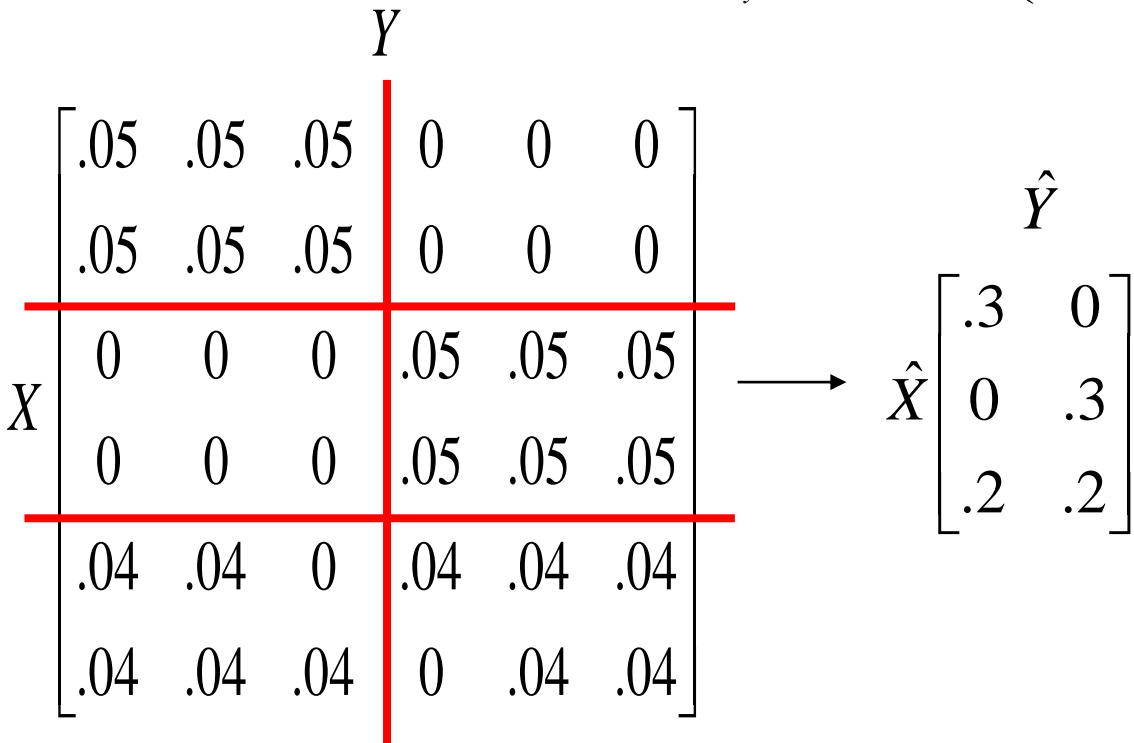


# Step 2: Information-Theoretic co-clustering links and classes

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$



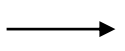
# Step 2: Information-Theoretic co-clustering links and classes

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

	Y					
X	.05	.05	.05	0	0	0
	.05	.05	.05	0	0	0
	0	0	0	.05	.05	.05
	0	0	0	.05	.05	.05
	.04	.04	0	.04	.04	.04
	.04	.04	.04	0	.04	.04



$$\hat{X} \hat{Y} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}$$

The optimal co-clustering is one that minimizes the difference (“loss”) in mutual information.

$$I(X,Y) - I(\hat{X},\hat{Y}) = .0957$$

## Step 2: Information-Theoretic co-clustering links and classes

- *ITCC + (Intra-similarity)*

$$I(X;Y) - I(\hat{X};\hat{Y}) - \lambda LCS - \mu CCS$$

$$LCS = \frac{1}{k} \sum_{i=1}^k \sum_{x,x' \in \hat{x}_i} \frac{sim(x, x')}{|\hat{x}_i| * (|\hat{x}_i| - 1)}$$

$$CCS = \frac{1}{l} \sum_{j=1}^l \sum_{y,y' \in \hat{y}_j} \frac{sim(y, y')}{|\hat{y}_j| * (|\hat{y}_j| - 1)}$$

- Compute the link similarity and class similarity by combining three measures(cosine, lexical and semantic similarity) based on a linear combination.

$$sim(c_i, c_j) = \alpha \cdot sim_{cos}(c_i, c_j) + \beta \cdot sim_{edit}(c_i, c_j) + \gamma \cdot sim_{sem}(c_i, c_j)$$



# Experiments

- Compared our approach with three baseline algorithms on real-world datasets.
- Implemented our approach in a prototype system, and then compared with two Linked Data browsers via a user study.

# Experimental Evaluation

## Dataset

- DBpedia1 *Mapping-based Properties* dataset, excluding RDF triples containing literals.
- 4 common classes (i.e., Artist, City, Company, University).  
For each class, we collected those entities that each one has more than 15 semantic links.
- As to entity classes, we used the *Mapping-based Types* dataset and the *DBpedia Ontology* dataset.

<http://wiki.dbpedia.org/Downloads2015-04>.

# Experimental Evaluation

## Dataset

Table 1. Statistics of experimental datasets

	Artist	City	Company	University
Number of entities	1233	2243	304	510
Number of links	139	280	174	163
Number of linked entities	59654	402580	88003	57487
Average num. of links per entity	25.8	30.1	27.2	28.9
Average num. of linked entities per entity	113.2	217.5	338.1	151.7
Average num. of classes per linked entity	6.9	8.7	5.8	8.9

# Experimental Evaluation

## Baselines

- Information-Theoretic Co-Clustering (**ITCC**), which not considers the intra-row and intra-column similarity.
- Bipartite spectral graph partition (**BSGP**, Dhillon 2001), which considers the co-clustering problem in term of finding minimum cut vertex partitions in a weighted bipartite graph.
- **K-means**. Since it is a one-sided clustering algorithm, the link and class collections are clustered separately.

# Experimental Evaluation

## Metrics

$$\textit{cohesion}(O) = \frac{1}{k} \sum_{i=1}^k \textit{coh}(O_i), \quad \textit{coh}(O_i) = \frac{\sum_{o \in O_i, o' \in O_i} \textit{sim}(o, o')}{|O_i| \cdot (|O_i| - 1)}.$$

$$\textit{separation}(O) = \frac{1}{k} \sum_{i=1}^k \textit{sep}(O_i), \quad \textit{sep}(O_i) = \frac{\sum_{o \in O_i, o' \notin O_i} \textit{sim}(o, o')}{|O_i| \cdot (N - |O_i|)}.$$

$$\textit{overall}(O) = 1 - \frac{\textit{separation}(O)}{\textit{cohesion}(O)}.$$

# Experimental Evaluation

## Processing

- Randomly selected 200 entities from our experimental dataset. For each selected entity, we conducted 10 runs using four algorithms (ITCC+, ITCC, BSGP and K-means) respectively and reported the average *overall*.
- For parameter settings, we investigated the sensitivity with respect to the disjoint clusters size  $k$  ( $=3, 5, 8$ ) and the balanced parameters of similarity computing ( $\alpha, \beta, \gamma$ ).

# Experimental Evaluation

## Results

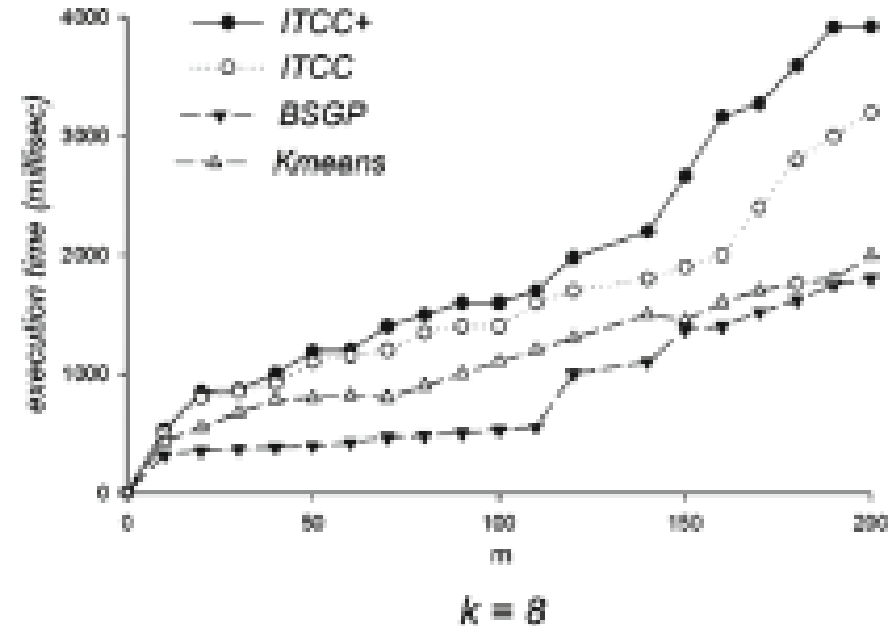
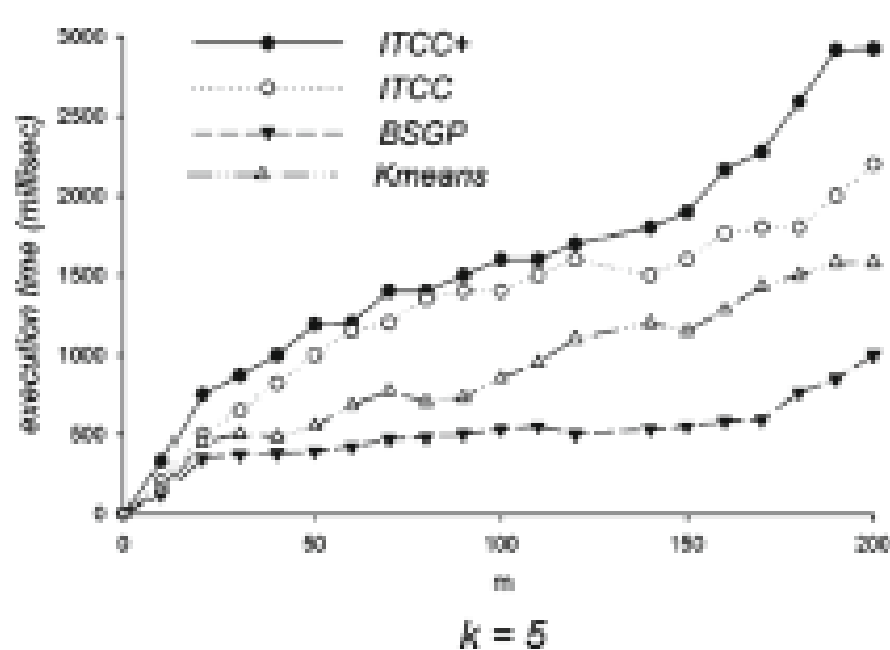
Comparison of *overall* against different  $k$  and  $(\alpha, \beta, \gamma)$

		$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$k=3$	ITCC+	<b>0.874</b>	0.566	0.401	<b>0.467</b>	<b>0.911</b>	0.642
	ITCC	0.872	0.377	<b>0.486</b>	0.273	0.688	<b>0.729</b>
	BSGP	0.835	0.226	0.179	0.238	0.465	0.43
	K-means	0.852	<b>0.643</b>	0.463	0.356	0.673	0.577
$k=5$	ITCC+	<b>0.891</b>	0.566	<b>0.587</b>	<b>0.428</b>	<b>0.925</b>	0.535
	ITCC	0.887	0.384	0.427	0.372	0.749	0.471
	BSGP	0.797	0.273	0.209	0.193	0.522	0.383
	K-means	0.814	<b>0.762</b>	0.479	0.383	0.765	<b>0.61</b>
$k=8$	ITCC+	<b>0.857</b>	<b>0.672</b>	<b>0.52</b>	0.433	<b>0.887</b>	0.668
	ITCC	0.832	0.359	0.497	0.324	0.886	<b>0.729</b>
	BSGP	0.686	0.238	0.419	0.273	0.596	0.365
	K-means	0.823	0.663	0.478	<b>0.481</b>	0.829	0.649



# Experimental Evaluation

## Efficiency Evaluation



Execution time with varying  $k$  and  $m$  (the size of current entities)

Time complexity of ITCC:  $O((nz(k + l) + km^2 + ln^2)\tau)$

# User Study

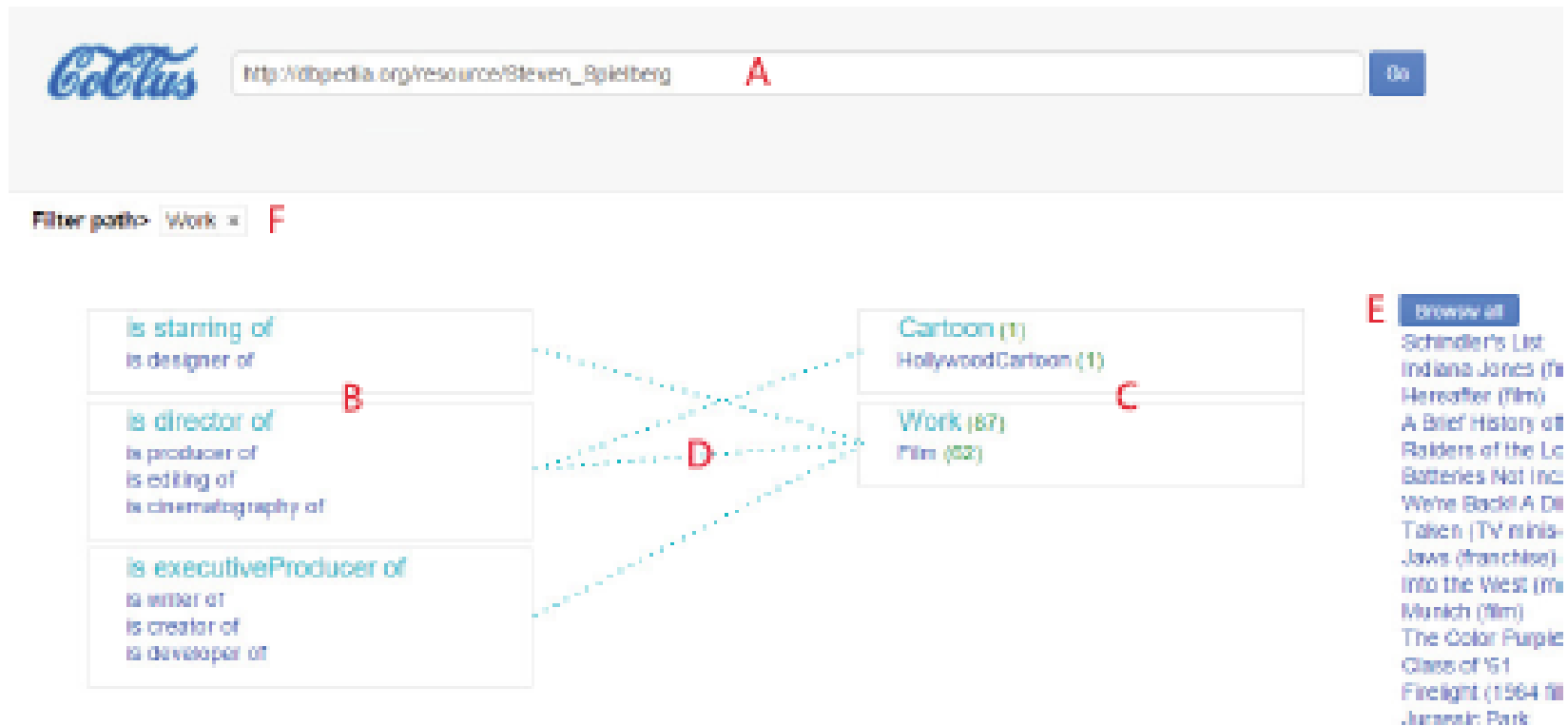
- 3 participant systems
  - CoClus
  - SView
  - Rhizomer
- 12 subjects
- 32 navigation tasks over DBpedia

## Navigation tasks about John Lennon

		Tasks
<i>G1</i>	<i>E1</i>	Explore the information related to John Lennon, and describe three main aspects of him
	<i>F1</i>	Find the albums written by John Lennon
<i>G2</i>	<i>E2</i>	Explore the information related to the band members of the Beatles, and describe three main aspects of them
	<i>F2</i>	Find the films starred by the band members of the Beatles

# User Study

- Overview of Prototype

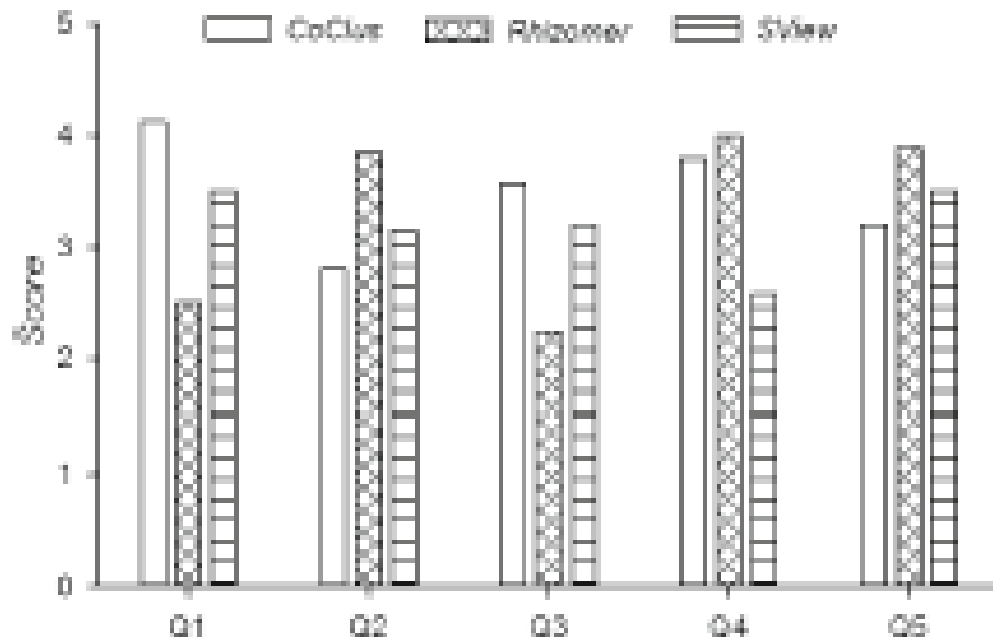


<http://ws.nju.edu.cn/coclus/>

# User Study

- Navigation Questionnaire

	Questions
Q1:	The system helped me get an overview of all the information
Q2:	The number of navigation options was overwhelming
Q3:	The navigation options were well organized
Q4:	The navigation option titles were understood well
Q5:	It was easy to reorient myself in the navigation



Results of  
navigation questionnaire.

# Conclusion

- We propose a co-clustering approach which clusters both links and classes simultaneously to provide users with an iterative entity navigation.
- We also measure the link similarity and the class similarity, and incorporate them into the co-clustering algorithm.
- The proposed approach is implemented in a prototype system. The evaluation results demonstrate that it supports users' iterative entity navigation.

THANK YOU!

QUESTIONS?