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

Liang Zheng, Jiang Xu, Jidong Jiang, Yuzhong Qu, Gong Cheng

Nanjing University, China

南京大学
NANJING UNIVERSITY

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* [e 'laða] (listen)), officially the **Hellenic Republic** (Greek: Ελληνική Δημοκρατία *Ellīnikḯ Dīmokratía* [elini'ci ðimokra 'ti.a]), also known since ancient times as **Hellas** (Ancient Greek: Ἑλλάς *Ellás*

Hellenic Republic
Ελληνική Δημοκρατία
Ellīnikḯ Dīmokratía (Greek)


Flag


Coat of arms

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;[1] Modern Greek: Ακρόπολη Αθηνών *Akrópoli Athinón*) 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

UNESCO World Heritage Site

Acropolis, Athens

Name as inscribed on the World Heritage List

Location Greece

Athens

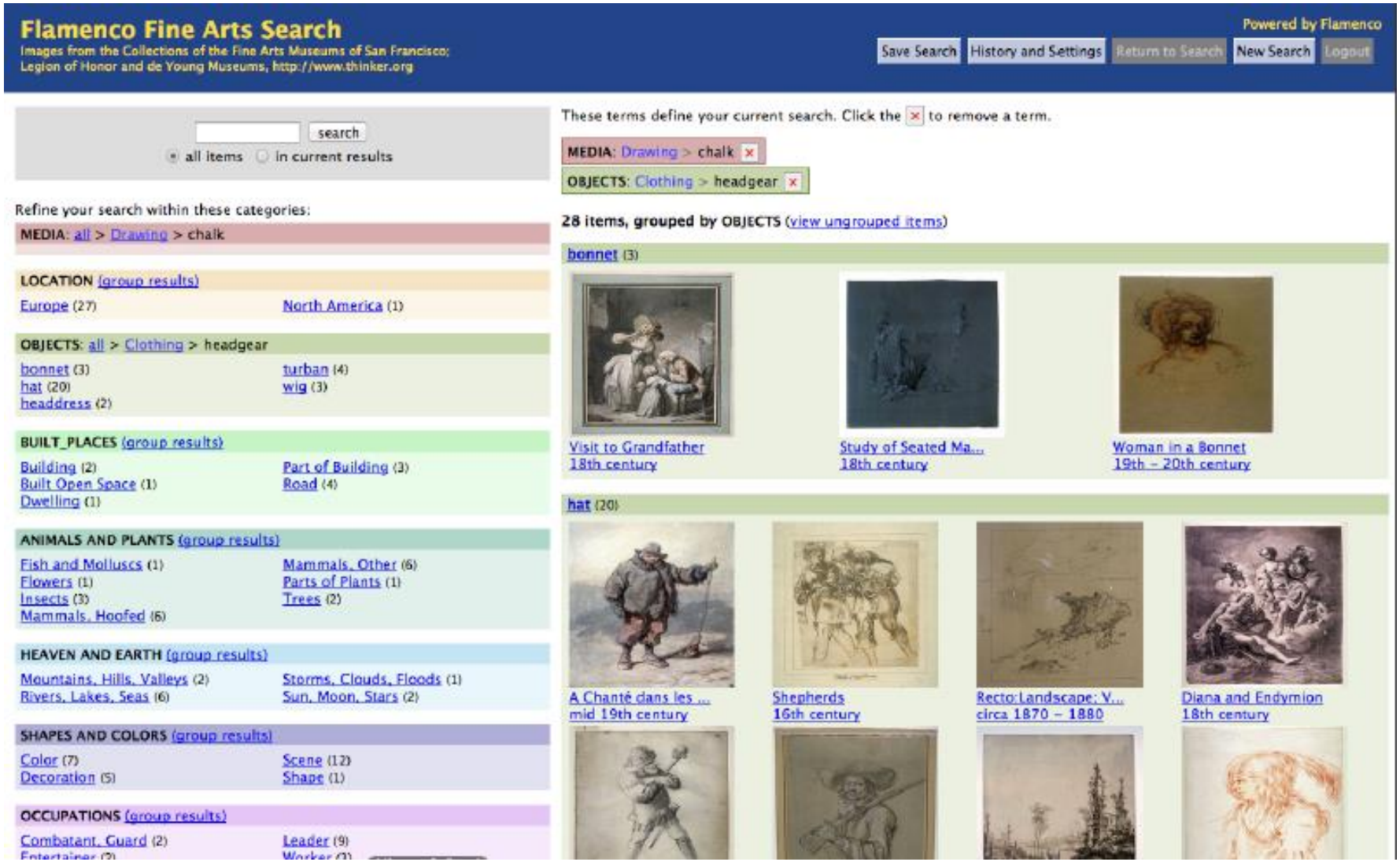
From Wikipedia, the free encyclopedia — Coordinates: 37° 58′ N 23° 43′ E

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

Athens (/'æθɪnz/;[2] Modern Greek: Αθήνα, *Athína* Greek pronunciation: [a'θina], 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.[3] Classical Athens was a powerful

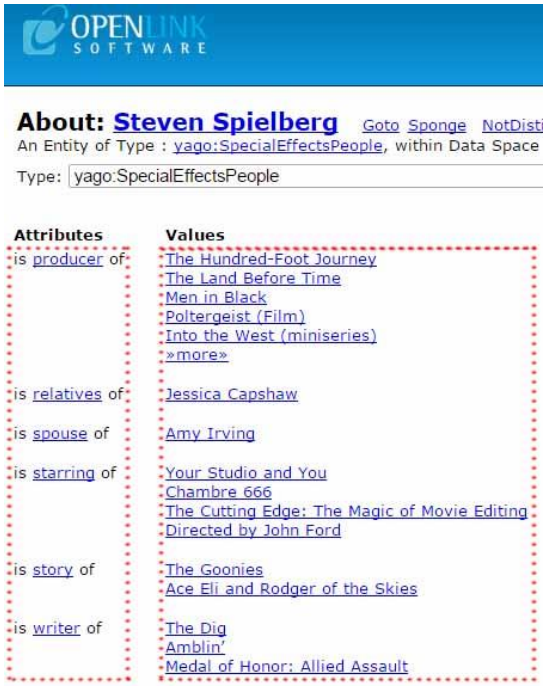
Athens
Αθήνα

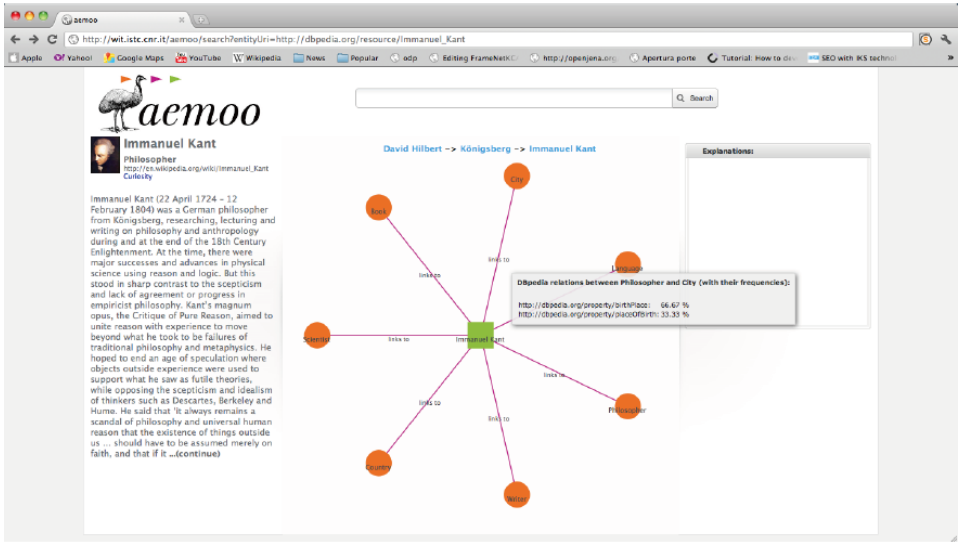
Faceted navigation



Entity Navigation

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





http://dbpedia.org/fct/

Aemoo (Alberto Musetti et al 2011)

Grouping related entities by type

Challenges:

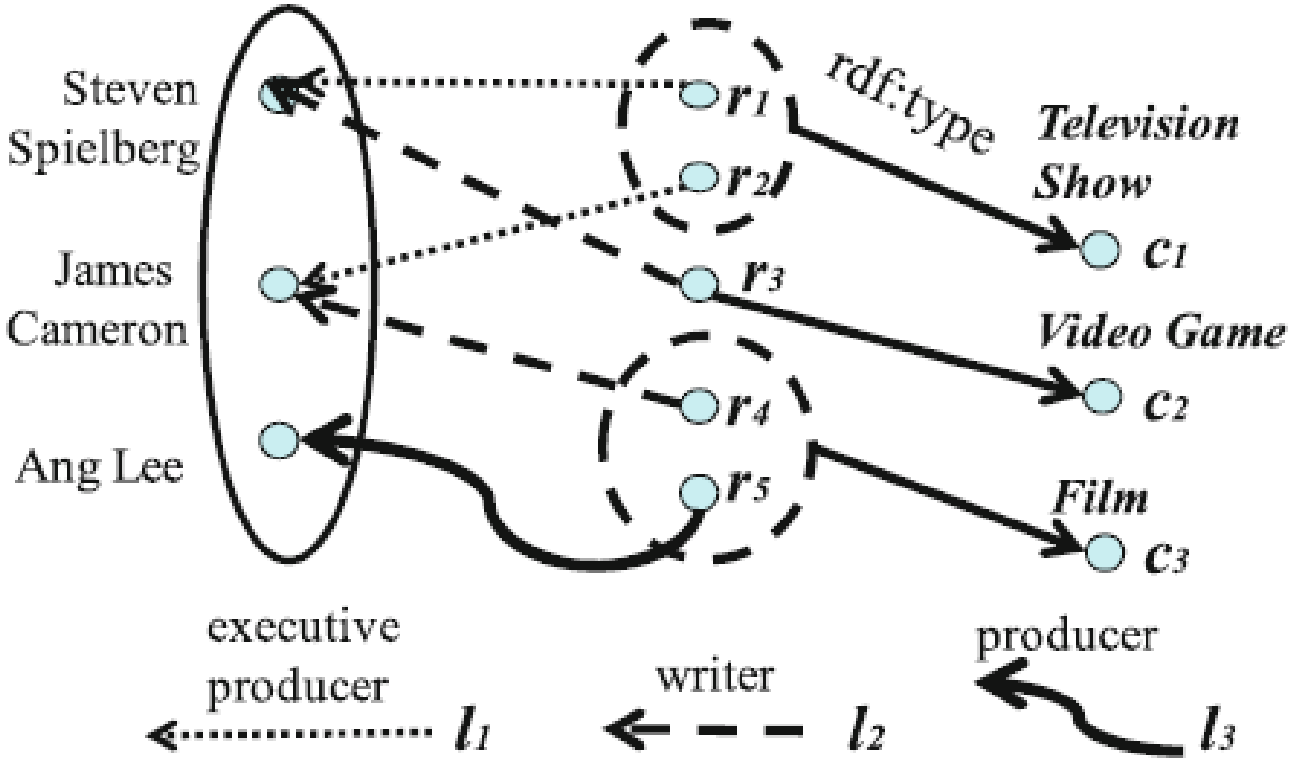
- **Large numbers** of linked entities and **high diversity** of links among entities, often make it hard for users to explore and find the entities of interest quickly.

Problem :

- How to improve the efficiency of entity navigation?

Motivation

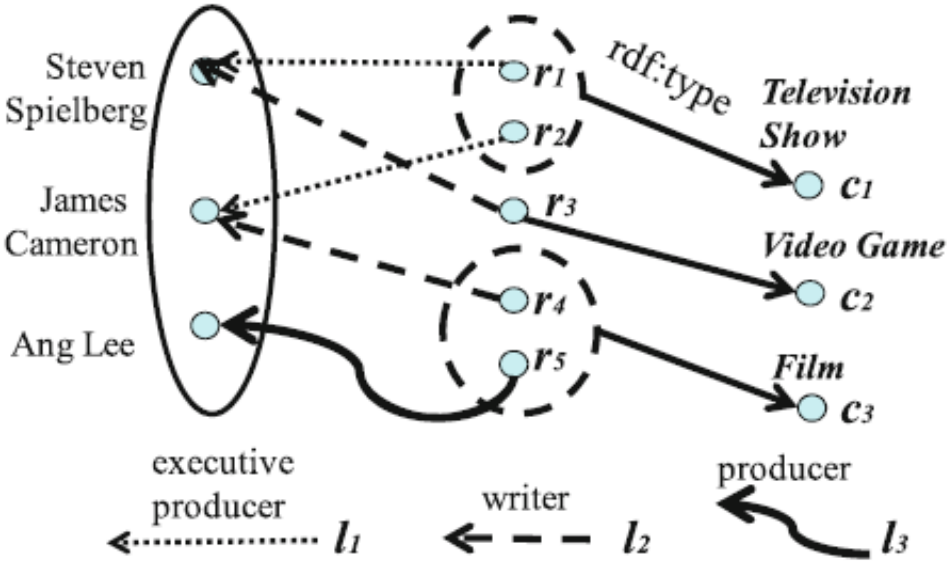
- As semantic links and classes of linked entities are two key aspects to help users navigation, clustering links and classes can offer effective ways to navigate over RDF data.

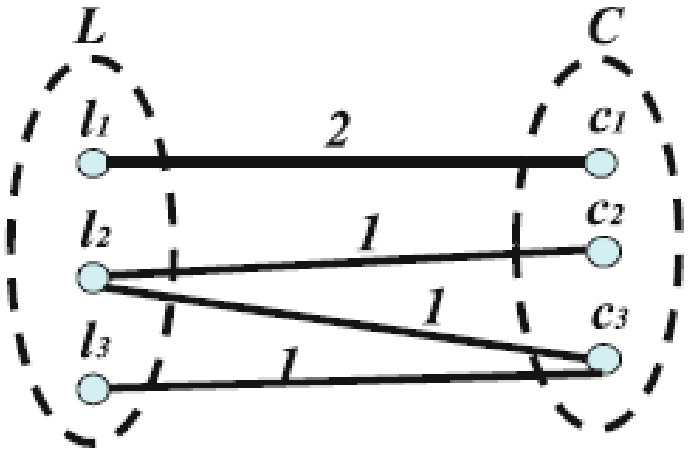


Approach

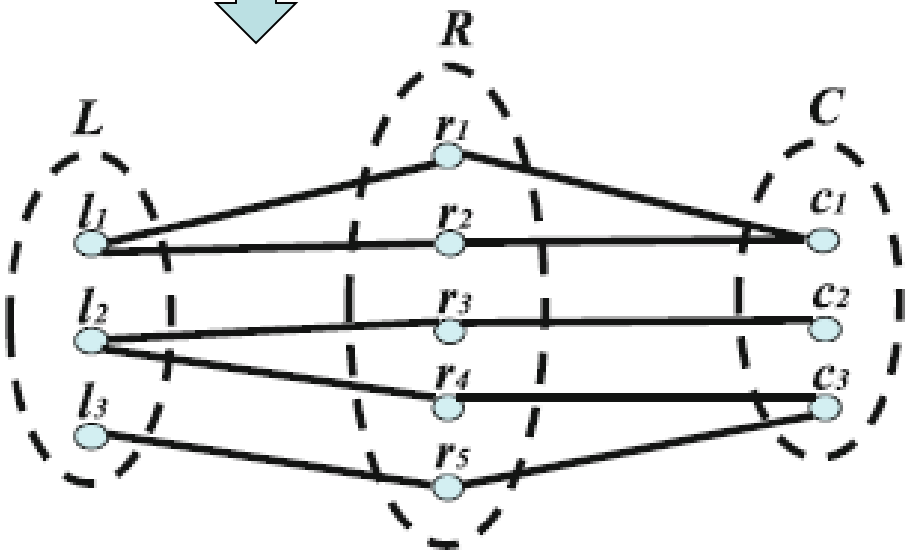
- We propose a co-clustering approach to provide users with iterative entity navigation.
- It clusters both links and classes simultaneously utilizing both the relationship between link and class, and the intra-link relationship and intra-class relationship.

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:

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

$$\begin{array}{c} \\ X \end{array} \begin{array}{c} \begin{array}{cccccc} & & & Y & & \end{array} \\ \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \end{array}$$

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

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

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

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

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

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

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

$$X \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \longrightarrow \hat{X} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}$$

Y $\qquad$ $\hat{Y}$

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

ITCC (Dhillon et al 2003)

- Mutual Information between random variables X and Y:

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

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

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

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

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

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

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

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 (α, β, γ).

Experimental Evaluation

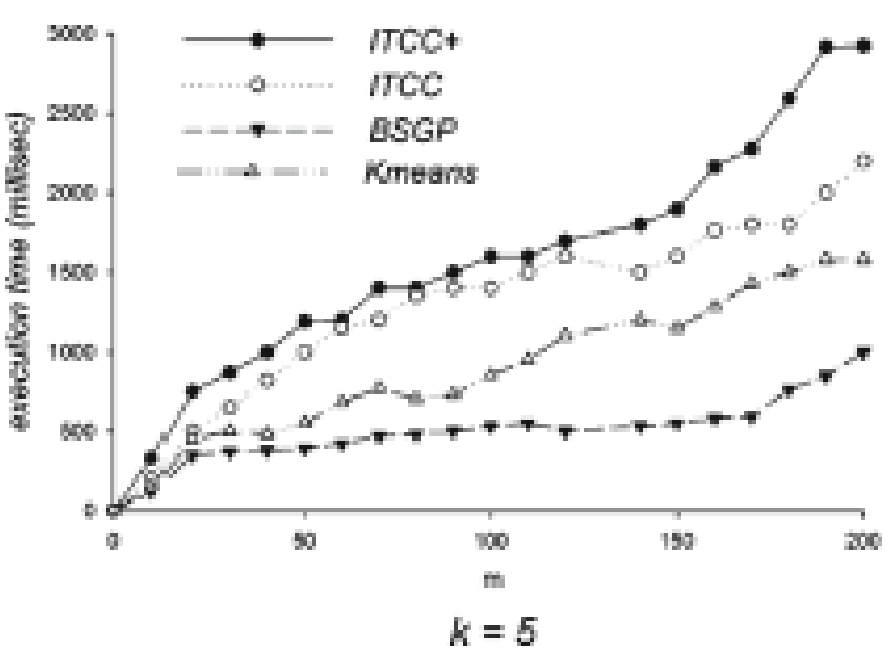
Results

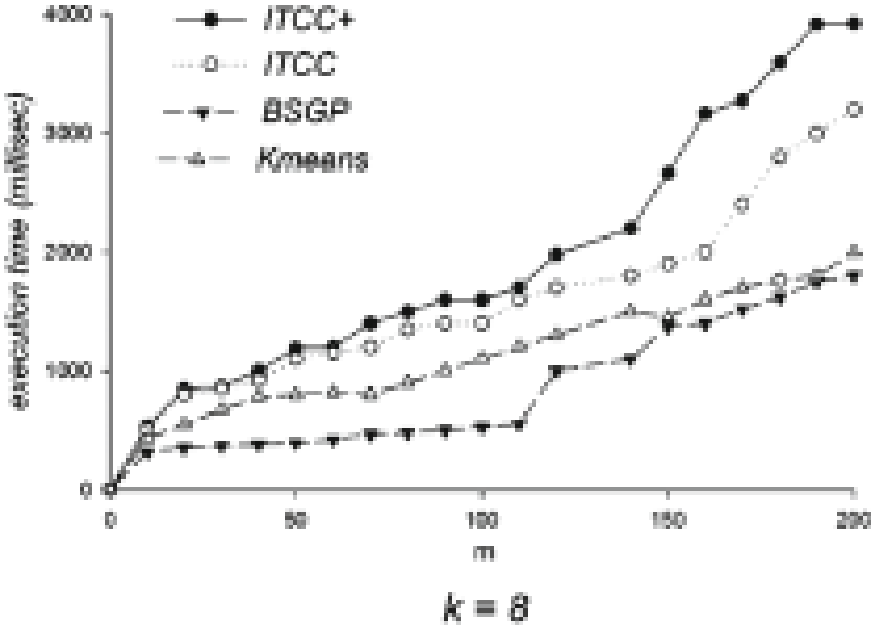
Comparison of *overall against different k and (α, β, γ)*

		t_1	t_2	t_3	t_4	t_5	t_6
k=3	ITCC+	**0.874**	0.566	0.401	**0.467**	**0.911**	0.642
	ITCC	0.872	0.377	**0.486**	0.273	0.688	**0.729**
	BSGP	0.835	0.226	0.179	0.238	0.465	0.43
	K-means	0.852	**0.643**	0.463	0.356	0.673	0.577
k=5	ITCC+	**0.891**	0.566	**0.587**	**0.428**	**0.925**	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	**0.762**	0.479	0.383	0.765	**0.61**
k=8	ITCC+	**0.857**	**0.672**	**0.52**	0.433	**0.887**	0.668
	ITCC	0.832	0.359	0.497	0.324	0.886	**0.729**
	BSGP	0.686	0.238	0.419	0.273	0.596	0.365
	K-means	0.823	0.663	0.478	**0.481**	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

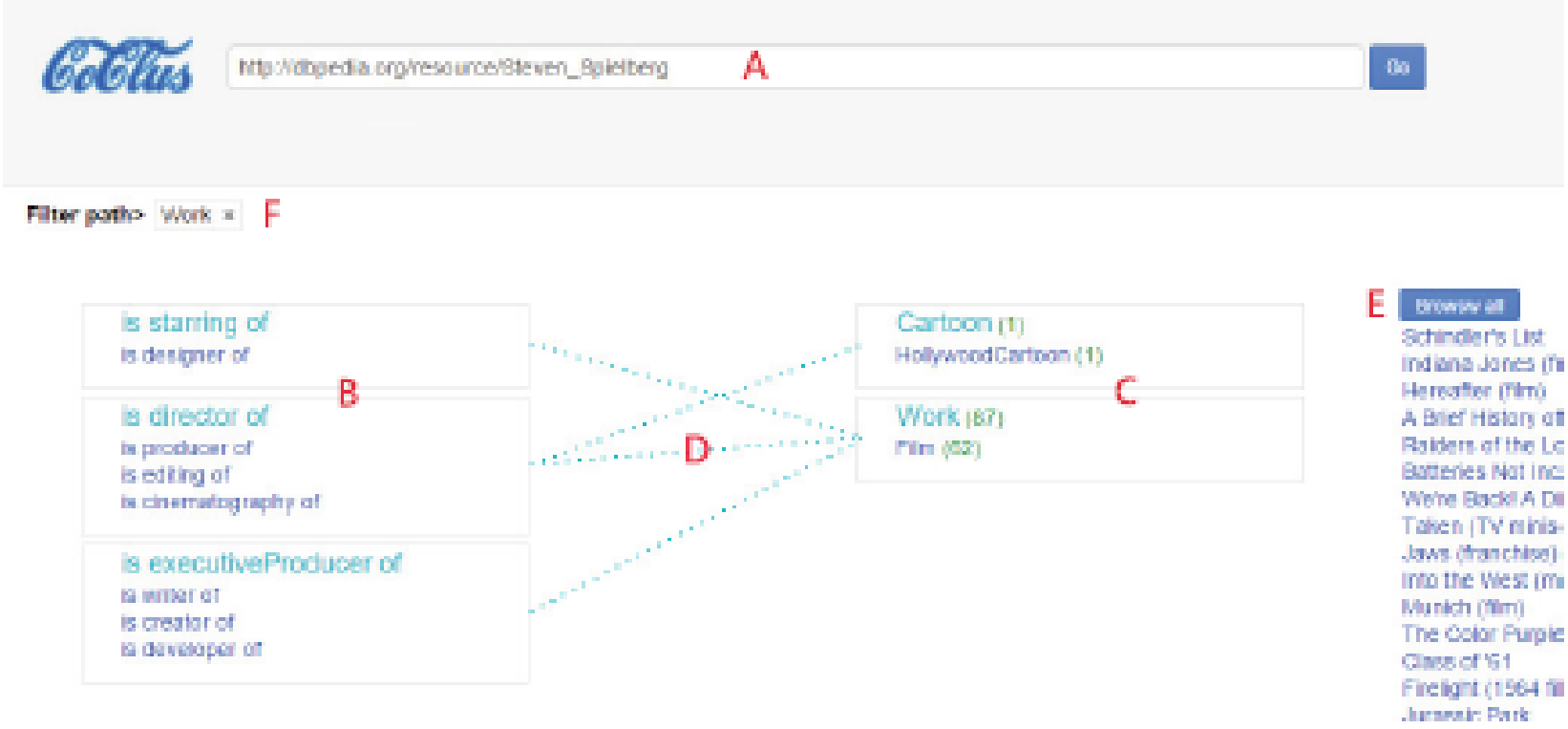
- 3 participant systems
  - CoClus
  - SView
  - Rhizomer
- 12 subjects
- 32 navigation tasks over DBpedia

Navigation tasks about John Lennon

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

User Study

- **Overview of Prototype**

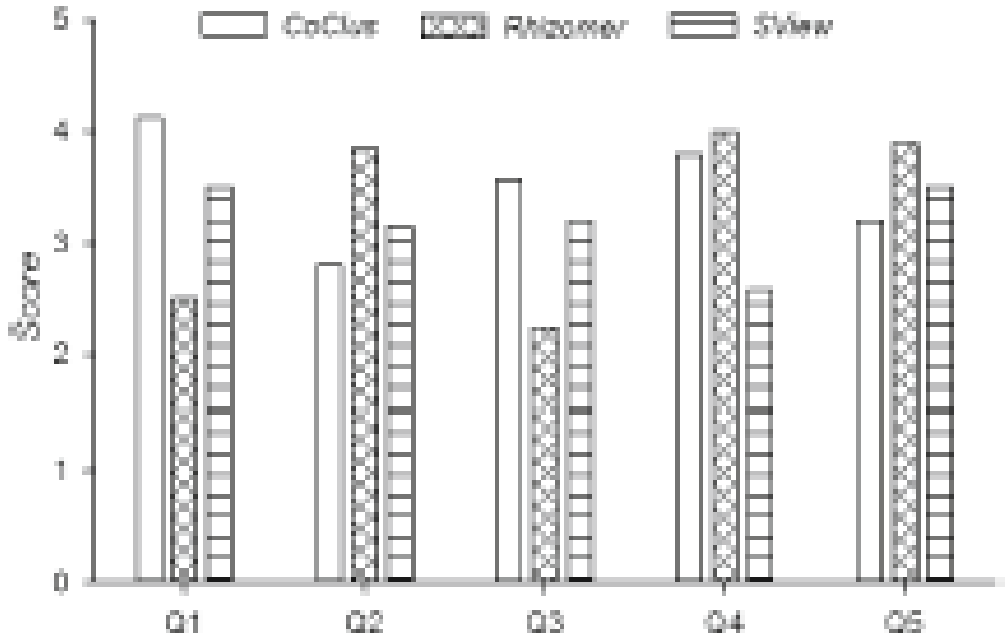


http://ws.nju.edu.cn/coclus/

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?