

Domain Specific Cross-lingual Knowledge Linking based on Similarity Flooding

Authors: Liangming PAN, Zhigang WANG, Juanzi LI, Jie TANG
Knowledge Engineering Group
Tsinghua University

2016-08-26



Outline

Backgrounds

Problem Definition

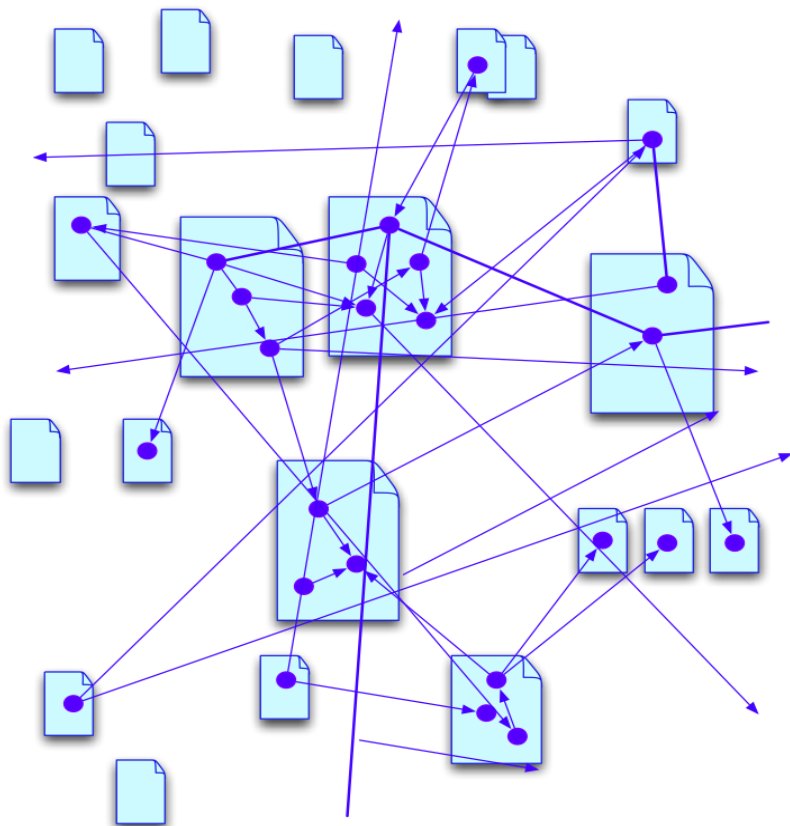
Methods

Experiments and Analysis

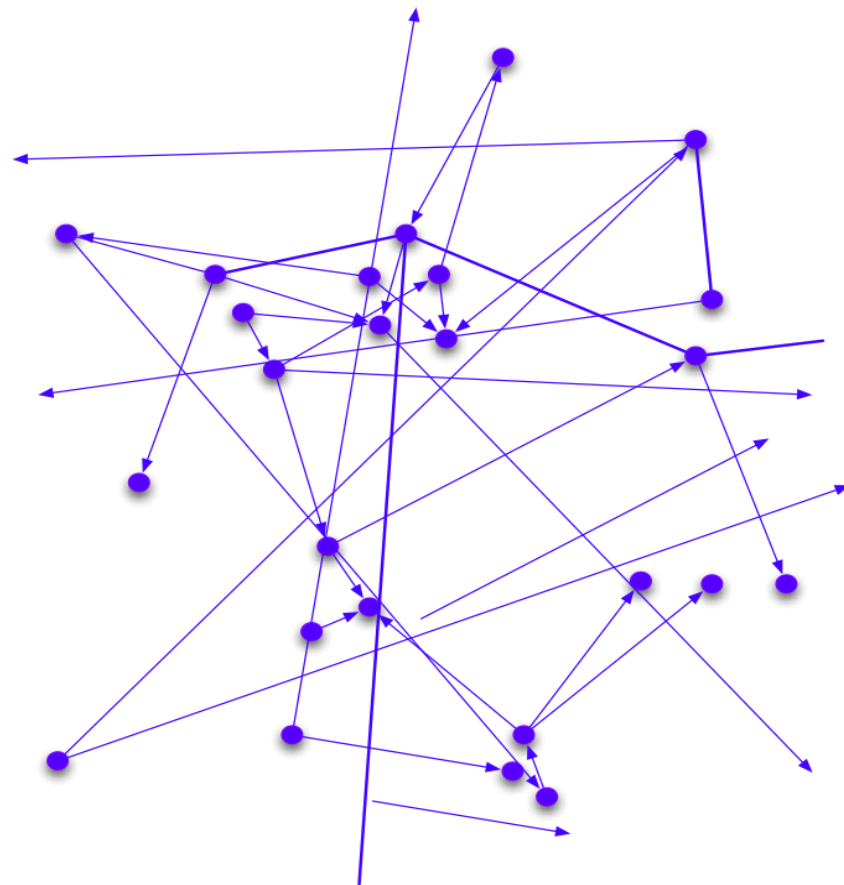
Conclusion

Semantic Web

World Wide Web

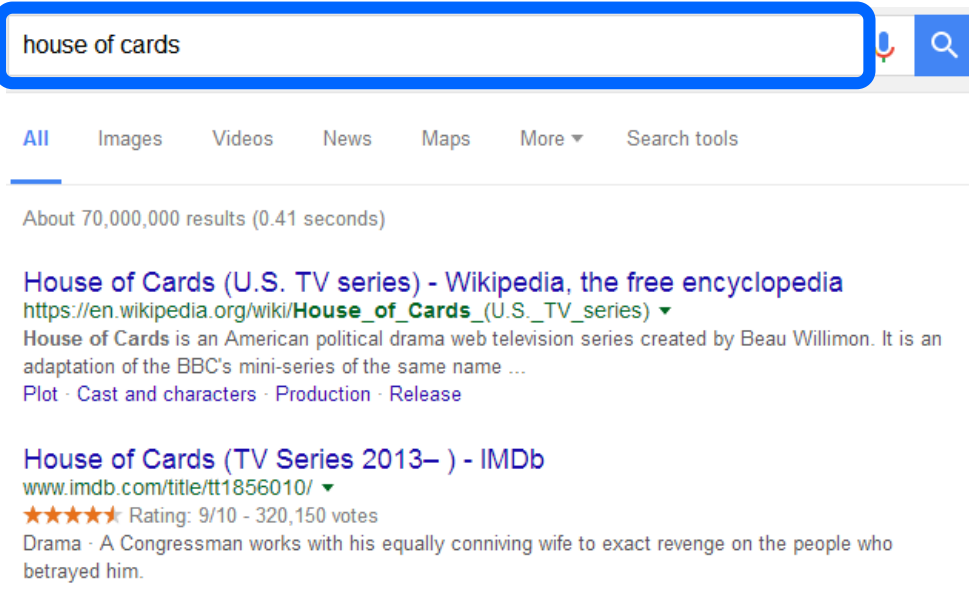


Semantic Web



Application: Web Search

Document-based ranking

house of cards

All Images Videos News Maps More Search tools

About 70,000,000 results (0.41 seconds)

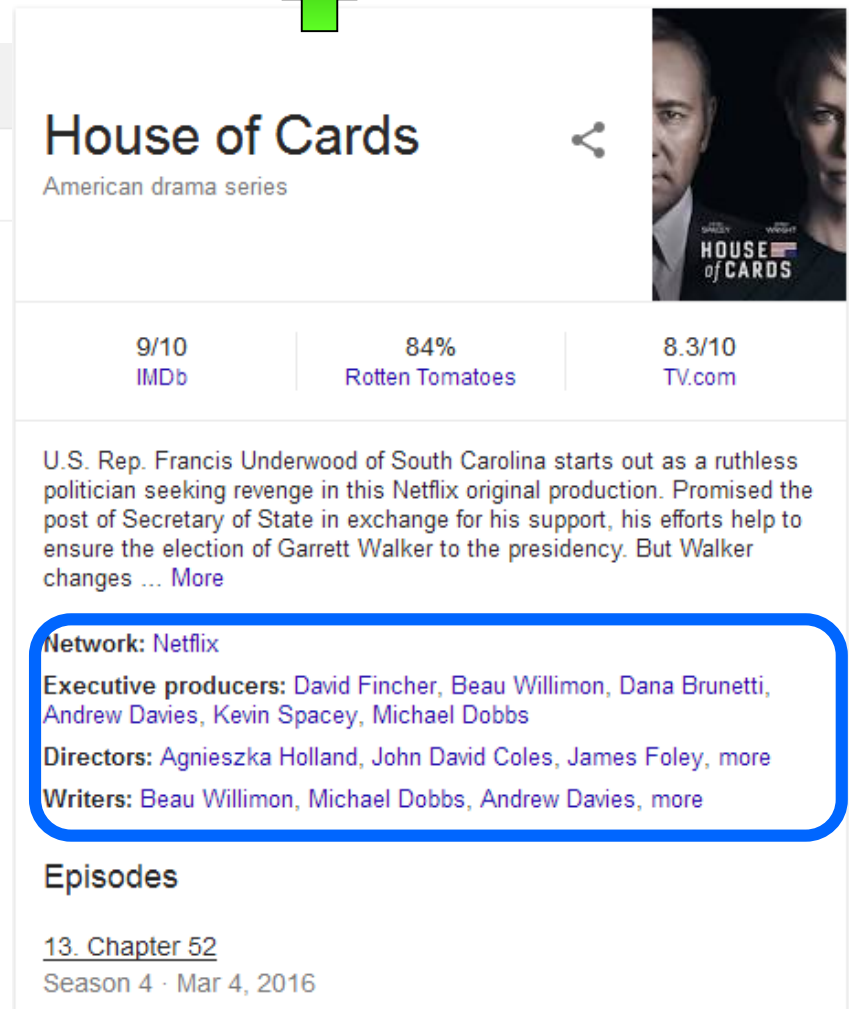
House of Cards (U.S. TV series) - Wikipedia, the free encyclopedia
[https://en.wikipedia.org/wiki/House_of_Cards_\(U.S._TV_series\)](https://en.wikipedia.org/wiki/House_of_Cards_(U.S._TV_series))

House of Cards is an American political drama web television series created by Beau Willimon. It is an adaptation of the BBC's mini-series of the same name ...
 Plot · Cast and characters · Production · Release

House of Cards (TV Series 2013–) - IMDb
www.imdb.com/title/tt1856010/

★★★★★ Rating: 9/10 - 320,150 votes
 Drama · A Congressman works with his equally conniving wife to exact revenge on the people who betrayed him.

Entity and relation summarization

House of Cards
 American drama series

9/10 IMDb	84% Rotten Tomatoes	8.3/10 TV.com
--------------	------------------------	------------------

U.S. Rep. Francis Underwood of South Carolina starts out as a ruthless politician seeking revenge in this Netflix original production. Promised the post of Secretary of State in exchange for his support, his efforts help to ensure the election of Garrett Walker to the presidency. But Walker changes ... [More](#)

Network: Netflix

Executive producers: David Fincher, Beau Willimon, Dana Brunetti, Andrew Davies, Kevin Spacey, Michael Dobbs

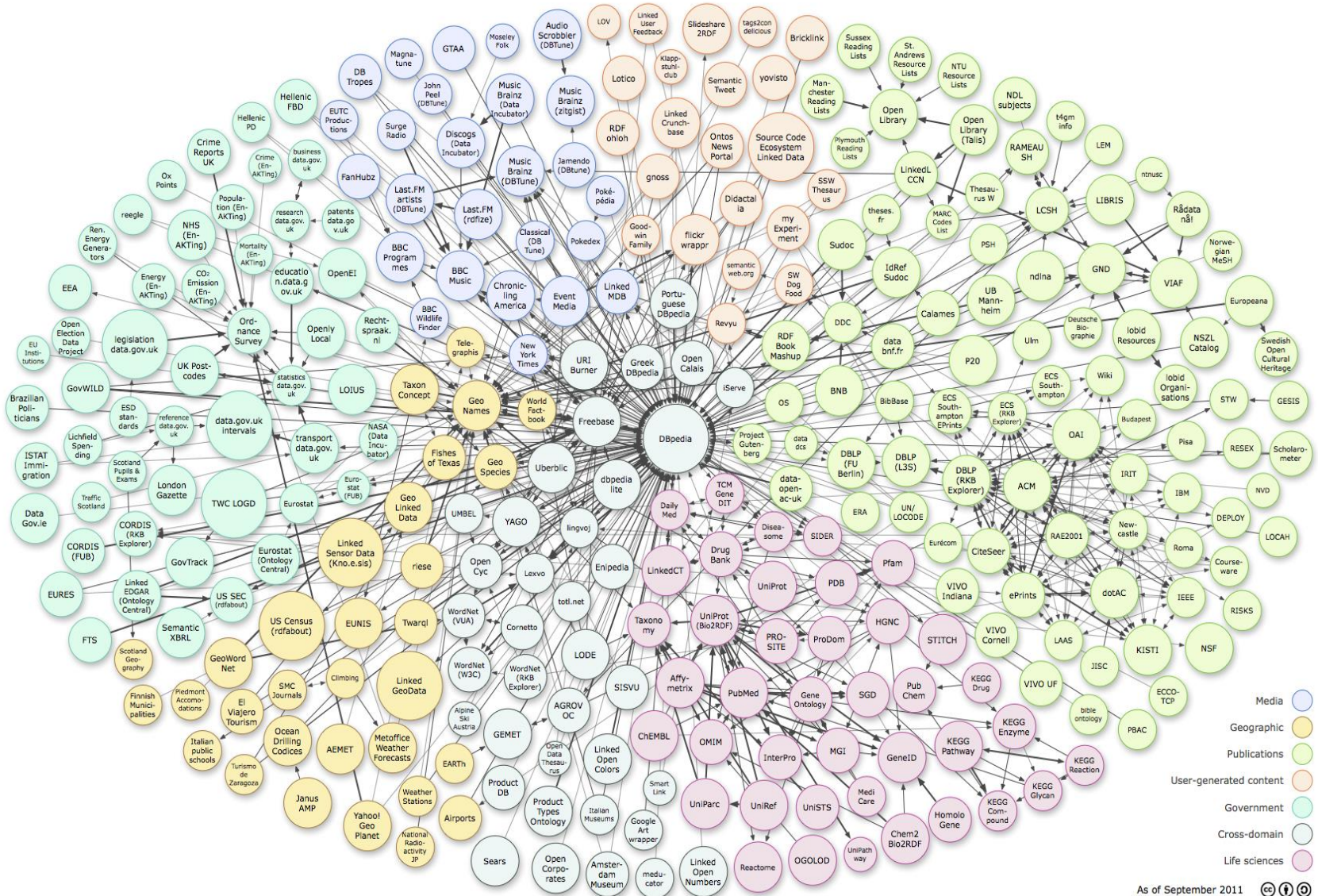
Directors: Agnieszka Holland, John David Coles, James Foley, more

Writers: Beau Willimon, Michael Dobbs, Andrew Davies, more

Episodes

13. Chapter 52
 Season 4 · Mar 4, 2016

Linked Open Data (LOD) Cloud



As of September 2011 

Online Encyclopedias

□ Hudong Baike

- 14,475,315 articles
- 10,605,432 contribution users



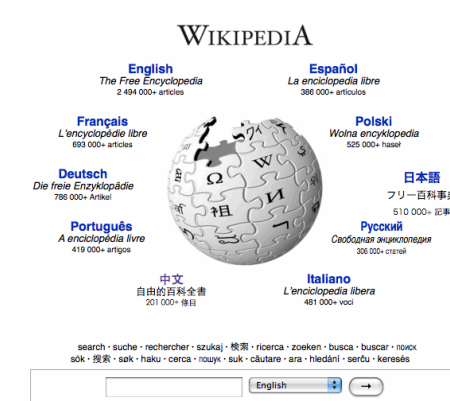
□ Baidu Baike

- 13,318,830 articles
- 5,793,900 contribution users



□ Wikipedia

- 863,918 articles in Chinese
- 5,138,426 articles in English



Cross-lingual Knowledge Linking

Link equivalent entities in different languages

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox
Print/export

Languages
Català
Deutsch
Español
Italiano
עברית
Nederlands
日本語

Article Discussion Read Edit View History Search

Anaerobic exercise Title

From Wikipedia, the free encyclopedia

Anaerobic exercise is exercise intense enough to trigger anaerobic metabolism. It is used by athletes in non-endurance sports to promote strength, speed and mass, and by bodybuilders to increase muscle mass. Muscles used in anaerobic exercise develop differently compared to aerobic exercise, leading to greater performance in short duration, high intensity activities, which last from mere seconds up to about 2 minutes.^{[1][2]} Any activity after about

Outlinks

EXERCISE ZONES

BEATS PER MINUTE	Very low intensity activity
	Anaerobic (hardcore training)
	Aerobic (Cardio training / Endurance)
	Weight control (Fitness / Fat burn)
Moderate activity (Maintenance / Warm up)	

Fox and Haskell formula

Categories: Exercise physiology

Categories

(cur | prev) 18:37, 27 October 2011 Reprint123 (talk | contribs) m (7,723 bytes) (Reverted edits by 88.17.84.74 (talk) to last version by Nepenthes) (undo)

(cur | prev) 15:37, 27 October 2011 88.17.84.74 (talk) (7,888 bytes) (undo)

(cur | prev) 15:38, 19 October 2011 Nepenthes (talk | contribs) m (7,723 bytes) (Reverted edits by 75.149.89.145 (talk) to last revision by Fayeman (HGO)) (undo)

(cur | prev) 15:38, 19 October 2011 75.149.89.145 (talk) (5,441 bytes) (→ Lactate threshold (LT) (or lactate inflection point (LIP))) (undo)

(cur | prev) 18:25, 14 October 2011 Fayeman (talk | contribs) (7,723 bytes) (Reverted good faith edits by 62.1.121.38 (talk): Removal of relevant content. (TW)) (undo)

Authors

Baidu 百科 无氧运动

首页 自然 文化 地理 历史 生活 社会 艺术 人物 经济 科技 体育 百科

无氧运动 Title

百科名片

无氧运动是指肌肉在“缺氧”的状态下高速剧烈的运动。无氧运动大部分是负荷强度高、瞬间性强的运动，所以很难持续长时间，而且疲劳消除的时间也慢。

Outlinks

无氧运动是相对有氧运动而言的。在运动过程中，身体的新陈代谢是加速的，加速的代谢需要消耗更多的能量。人体的能量是通过身体内的糖、蛋白质和脂肪分解代谢得来的。在运动量不大时，比如慢跑、打羽毛球、跳舞等情况下，机体能量的供应主要来源于脂肪的有氧代谢，以脂肪的有氧代谢为主要供应能量的运动就是我们所说的有氧运动。当我们从事的运动非常剧烈，或者是急速爆发的，例如举重、百米冲刺、摔跤等，此时机体在瞬间需要大量的能量，而在正常情况下，有氧代谢是不能满足身体此时的需求的，于是糖就进行无氧代谢，以快速产生大量能量。这种状态下的运动就是无氧运动。

特征

无氧运动的最大特征是：运动时氧气的摄入量非常低。由于速度过快及爆发力过猛，人体内的糖分来不及经过氧气分解，而不得不依靠“无氧供能”。这种运动会在体内产生过多的乳酸，导致肌肉疲劳不能持久，运动后感到肌肉酸痛，呼吸急促。其实是发酵时产生大量丙酮酸、乳酸等中间代谢产物，不能通过呼吸排除。这些酸性产物堆积在细胞和血液中，就成了“疲劳毒素”，会让人感到疲乏无力、肌肉酸痛，还会出现呼吸、心跳加快和心律失常，严重时会出现酸中毒和增加肝肾负担。所以无氧运动后，人总会疲惫不堪，肌肉疼痛要

合作编辑者

家用健身器械, 1316970089, happy快乐1122, 百科ROBOT, 汪惠er, lge200758

如果您认为本词条还需要进一步完善，百科欢迎您一起来参与 编辑词条 在开始编辑前，您还可以先学习如何编辑词条

Authors

开放分类：

运动, 无氧运动, 人体运动分类, 无氧肌肉训练

Categories

Cross-lingual links



Outline

Backgrounds

Problem Definition

Methods

Experiments and Analysis

Conclusion

Cross-lingual Knowledge Linking

- Given two Online Encyclopedias, K and K' , a *correspondence* between entities $e \in K$ and $e' \in K'$, denoted as $\langle e, e' \rangle$, signifies that e and e' are equivalent.
- Given two wiki knowledge bases, K and K' , and an initial set of correspondences $A = \{\langle e, e' \rangle_j\}_{j=1}^m$, *knowledge linking* is the process of finding more correspondences between K and K' .
- If K and K' are in different languages, we call it the problem of *cross-lingual knowledge linking*.

Related Works

Arthors	Task
Sorg and Cimiano 2008	Infer new CLs between German Wikipedia and English Wikipedia
Erdmann et al. 2009	Extracted a dictionary from Wikipedia by analyzing the link structure of Wikipedia.
Hassan et al. 2009	Address the task of cross-lingual semantic relatedness
Wang et al. 2012	Find CLs between English Wikipedia and Chinese Wikipedia.

Motivation

Motivation 1. The limited coverage of existing cross-lingual links

- Large number of known CLs are required in existed methods for cross-lingual knowledge linking for serving as either training data or seed set.
- There are less existing CLs or none at all between different wikis.

Motivation 2. Domain-Specific features are not fully utilized

- The existing methods have been entirely focus on the general framework for knowledge linking, with less focus on finding CLs in specific domains.
- Semantics carried by the properties in the infoboxes of wiki can be utilized when we focus on specific domains.

Domain Specific Cross-lingual Knowledge Linking

- If K and K' are from a specific domain, we call it the problem of *domain-specific cross-lingual knowledge linking*.

title → Steve Jobs

From Wikipedia, the free encyclopedia

Steven Paul "Steve" Jobs (/dʒɒbz/; February 24, 1955 – October 5, 2011) was an American information technology entrepreneur and inventor. He was the co-founder, chairman, and chief executive officer (CEO) of **Apple Inc.**; CEO and majority shareholder of **Pixar Animation Studios**; [3] a member of The **Walt Disney Company**'s board of directors following its acquisition of Pixar; and founder, chairman, and CEO of **NeXT Inc.** Jobs is widely recognized as a pioneer of the **microcomputer revolution** of the 1970s and 1980s, along with Apple co-founder Steve Wozniak. Shortly after his death, Jobs's official biographer, **Walter Isaacson**, described him as a "creative entrepreneur whose passion for perfection and ferocious drive revolutionized six industries: personal computers, animated movies, music, phones, tablet computing, and digital publishing." [2]

link → Apple Inc., Pixar Animation Studios, Walt Disney Company, NeXT Inc., microcomputer revolution

text → [Main paragraph]

info ←

Jobs in 2007	
Born	Steven Paul Jobs
attr	February 24, 1955 <i>value</i>
	San Francisco, California, U.S.
Died	October 5, 2011 (aged 56)
	Palo Alto, California, U.S.
Cause of death	Pancreatic cancer and respiratory arrest
Nationality	American
Ethnicity	German and Syrian

If we focus the problem in a given domain, *Infoboxes* can also be utilized for cross-lingual knowledge linking.



Outline

Backgrounds

Problem Definition

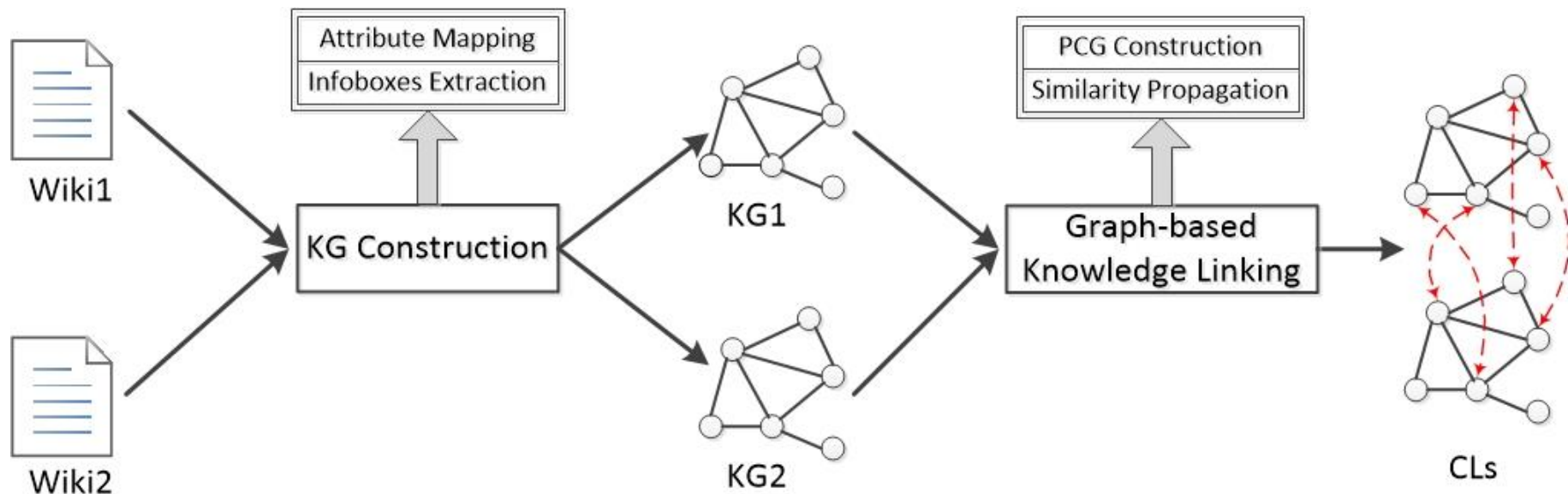
Methods

Experiments and Analysis

Conclusion

The Proposed Approach

General Framework



■ Step1: Knowledge Graph Construction

- Extract semantic relations in a structured form of subject-predicate-object triples from infoboxes of the two input wikis to construct two knowledge graphs.

■ Step2: Graph-based Knowledge Linking

- Discover CLs between the two constructed graphs based on a variation of the Similarity Flooding algorithm.

The Proposed Approach

Knowledge Graph Construction

Input 1: **Attribute Mapping** of a given domain

AM	s	s'
R_1	actor; starring; actorlist; starring list	主演; 演员; 演员表; 主演表
R_2	writer; screenwriter; writtenby; storyby; scenarist	编剧; 编剧列表; 剧本
R_3	director; directedBy; directing	导演; 导演人; 导演表
R_4	works; worklist; acting	作品; 作品列表; 代表作品; 主要作品; 参演作品

Input 2: **Wiki articles**

title → Steve Jobs

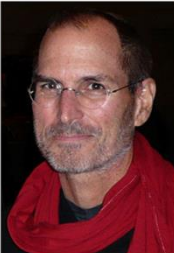
From Wikipedia, the free encyclopedia

Steven Paul "Steve" Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011) was an American information technology entrepreneur and inventor. He was the co-founder, chairman, and chief executive officer (CEO) of **Apple Inc.**, CEO and majority shareholder of Pixar Animation Studios,^[3] a member of The Walt Disney Company's board of directors following its acquisition of Pixar, and founder, chairman, and CEO of NeXT Inc. Jobs is widely recognized as a pioneer of the microcomputer revolution of the 1970s and 1980s, along with Apple co-founder Steve Wozniak. Shortly after his death, Jobs's official biographer, Walter Isaacson, described him as a "creative entrepreneur whose passion for perfection and ferocious drive revolutionized six industries: personal computers, animated movies, music, phones, tablet computing, and digital publishing."^[2]

link → **Apple Inc.**, **Pixar Animation Studios**, **Walt Disney Company**

text → drive revolutionized six industries: personal computers, animated movies, music, phones, tablet computing, and digital publishing."^[2]

Steve Jobs



Born Jobs in 2007

attr Steven Paul Jobs

value February 24, 1955

San Francisco, California, U.S.

Died October 5, 2011 (aged 56)

Palo Alto, California, U.S.

Cause of death Pancreatic cancer and respiratory arrest

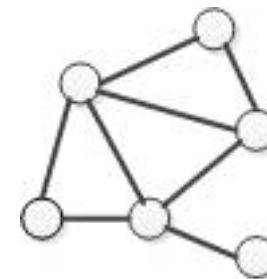
Nationality American

Ethnicity German and Syrian

info ←

→ Infoboxes

→ Inner Links



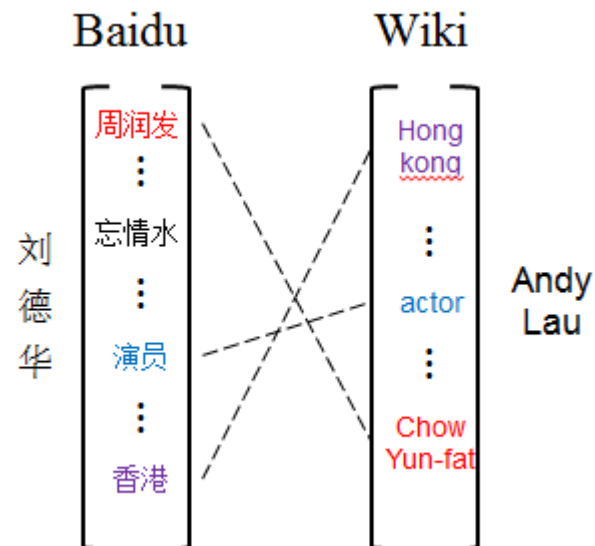
Knowledge Graph

The Proposed Approach

Graph-based Knowledge Linking

Step 1: Initial Similarity Computation

- Representing an entity as the vector of its text description.
- Incorporating a **domain dictionary** to provide translations for common domain terms. (e.g. Actor——演员)
- The initial similarity between two entities is the cosine similarity of their corresponding vectors.



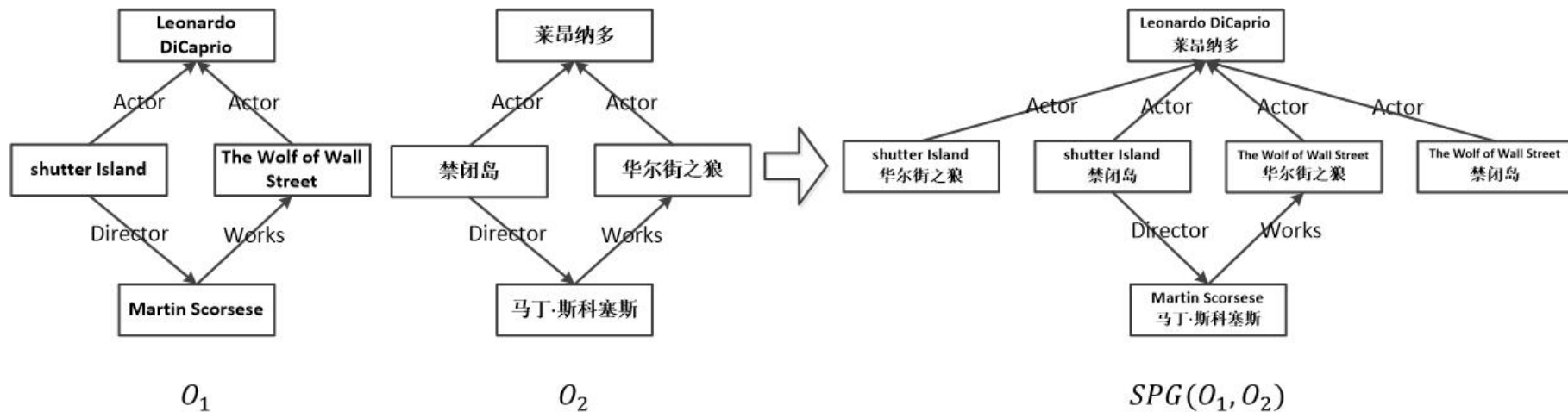
The Proposed Approach

Graph-based Knowledge Linking

Step 2: Propagation Graph Construction

Definition: Similarity Propagation Graph

- Given two knowledge graphs G and G' , $((e, e'), r, (o, o')) \in SPG(G, G')$ if and only if:
 - $(e, r, o) \in G$
 - $(e', r, o') \in G'$
 - $Sim(e, e') > \theta$
 - $Sim(o, o') > \theta$
- The $Sim(e, e')$ indicates the initial similarity between entity e and e' . θ is a pre-given threshold.



The Proposed Approach

□ Graph-based Knowledge Linking

■ Step 3: Similarity Flooding

$$\sigma^{i+1}(e, e') = \frac{1}{Z} (\sigma^0(e, e') + \sigma^i(e, e') + \varphi^i(e, e'))$$

$$\varphi^i(e, e') = \sum_{(o, o') \in IN(e, e')} \omega(o, o') \cdot \sigma^i(e, e')$$

$$Z = \max_{(e, e') \in SPG(G, G')} (\sigma^{i+1}(e, e'))$$

- $\varphi^i(e, e')$: the similarity gain from neighbors of (o, o') .
- $IN(e, e')$: the set of incoming neighbors of the node (e, e') in SPG.
- $\omega(e, e')$: the propagation weight which is simply defined as the inverse of the number of out-linking relationships for the node (o, o') .
- Z : the normalization factor.

Outline

Backgrounds

Problem Definition

Methods

Experiments and Analysis

Conclusion

Experiments and Analysis

□ Datasets

- 3 novel datasets extracted from wikis of movie domain
 - **EWM**: English Wikipedia of movie domain, 222,022 articles
 - **ZWM**: Chinese Wikipedia of movie domain, 112,164 articles
 - **BBM**: Baidu Baike of movie domain, 58,638 articles

Dataset	#Nodes	#Edges				
		<Actor>	<Director>	<Writer>	<Works>	<relatedTo>
EWM	220,989	185,453	73,705	48,500	373,550	681,208
ZWM	57,842	81,717	23,544	11,730	93,257	151,299
BBM	111,768	154,112	18,921	9,603	180,370	363,006

Table 1. Statistics of knowledge graphs for three datasets

■ Evaluation benchmarks

- 2,678 CLs between EWM and ZWM.
- 4,022 CLs between EWM and BBM.

Experiments and Analysis

□ Comparison Methods

■ Title Edit Distance (TED)

- This method simply translates the titles of Chinese articles into English by Google Translation API.

■ Initial Similarity (IS)

- This method directly regards the initial similarities as the final result.

■ Simple Similarity Propagation (SSP)

- This method conducts the similarity propagation process without the influence of initial similarity. We initialize all nodes in the SPG with a unified initial similarity.

■ Linkage Factor Graph (LFG)

- The LFG model is a state-of-art method for cross-lingual knowledge linking. The method first calculates several language-independent features from input wikis and then proposes a factor graph model to discover cross-lingual links.

Experiments and Analysis

Results Analysis

Tasks	Metrics	Methods				
		TED	IS	SSP	LFG	Proposed
EWM-BBM	$P@1$	55.32	68.14	77.61	83.73	89.89
	$P@5$	62.91	75.53	86.56	88.21	93.28
EWM-ZWM	$P@1$	54.79	61.88	70.11	80.26	83.51
	$P@5$	61.53	67.03	80.35	82.29	87.33

Table 2. Performance of knowledge linking with different methods (%).

- TED: only the entity titles are used in this method.
- IS: only the article texts of entities are used in this method.
- SSP: only the semantic information contained in the infoboxes are used.
- LFG: using structural features but semantics of infoboxes are not used.
- The Proposed: jointly exploits both the texture features, structural information and semantic information of wiki articles.



Outline

Backgrounds

Problem Definition

Methods

Experiments and Analysis

Conclusion



Thanks!

Liangming Pan

KEG, THU

peterpan10211020@163.com