Domain Specific Cross-lingual Knowledge Linking based on Similarity Flooding

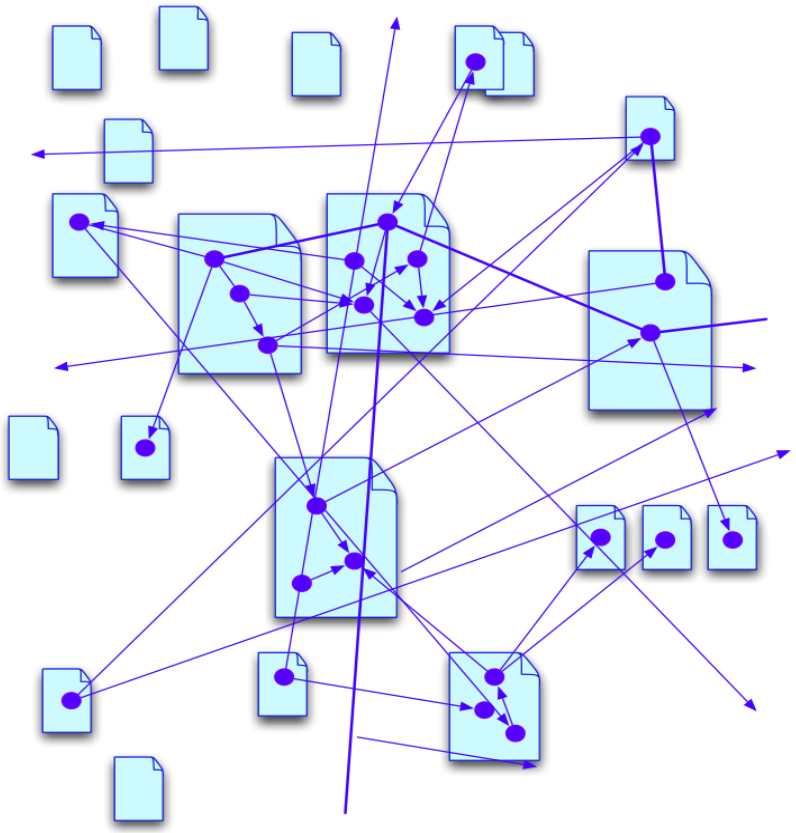
Authors: Liangming PAN, Zhigang WANG, Juanzi LI, Jie TANG
Knowledge Engineering Group
Tsinghua University

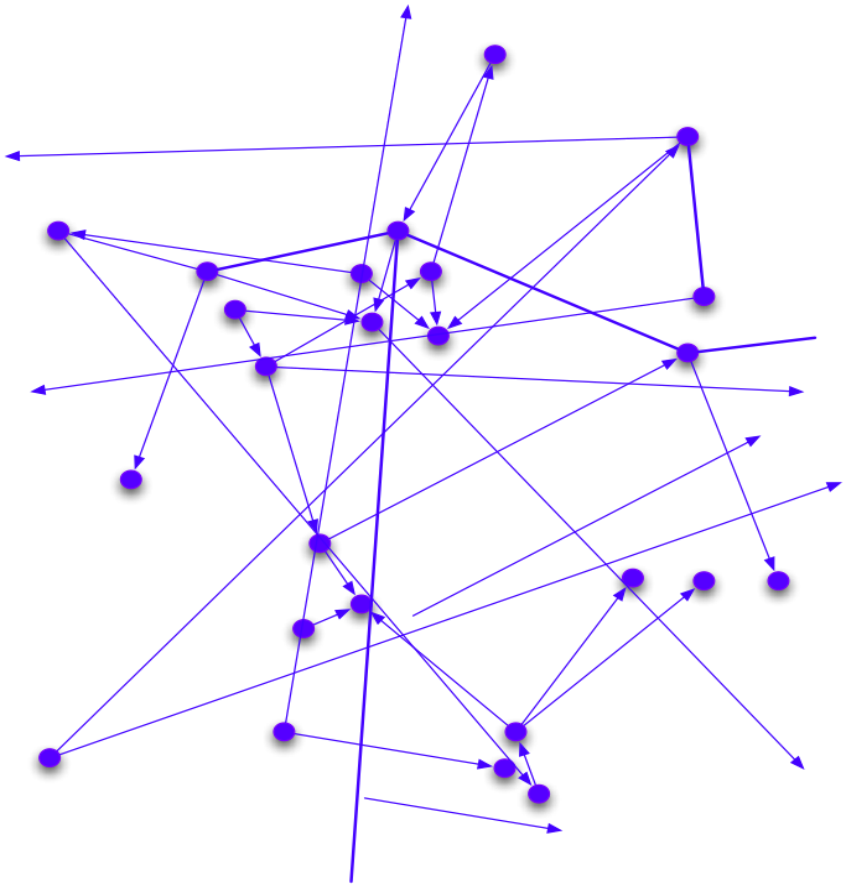
2016-08-26

Outline

- Backgrounds
- Problem Definition
- Methods
- Experiments and Analysis
- Conclusion

Semantic Web





Application: Web Search

Document-based ranking

Entity and relation summarization

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)
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/
★★★★½ Rating: 9/10 - 320,150 votes
Drama · A Congressman works with his equally conniving wife to exact revenge on the people who betrayed him.

In the news

'House Of Cards' Season 5 Air Date, Spoilers, News & Update: Mahershala Ali Leaves? Will Beau Willimon's Exit ...
Gamenguide - 2 days ago
Mahershala Ali will not reprise his role as Remy Danton in "**House of Cards**" Season 5 and ...

House of Cards

American drama series

9/10	84%	8.3/10
IMDb	Rotten Tomatoes	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

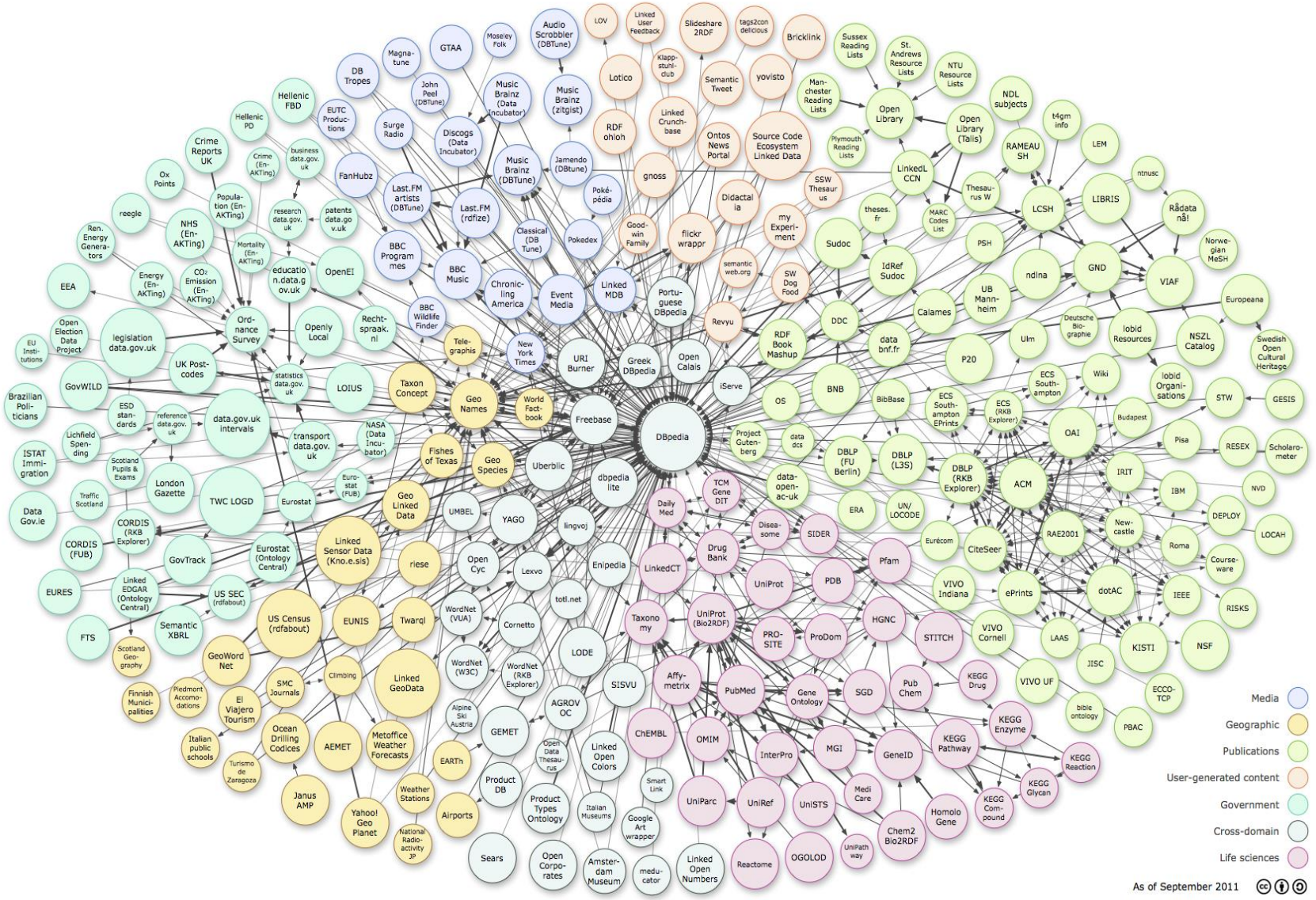
Directors: Agnieszka Holland, John David Coles, James Foley, more

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

Episodes

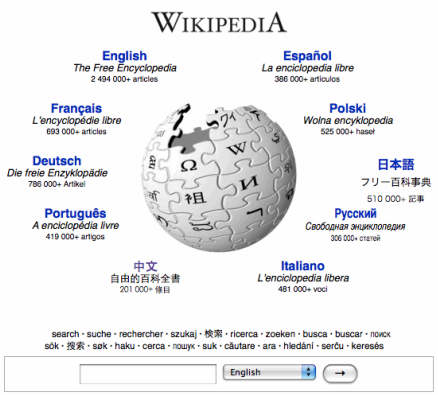
13. Chapter 52
Season 4 · Mar 4, 2016

Linked Open Data (LOD) Cloud



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

Outline

- Backgrounds
- Problem Definition
- Methods
- Experiments and Analysis
- Conclusion

Cross-lingual Knowledge Linking

- Given two Online Encyclopedias, K and K', a ***correspondence*** between entities $e \in \mathrm{K}$ and $e' \in \mathrm{K}'$, 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 $\mathrm{A} = \{\langle e, e' \rangle_j\}_{j=1}^{m}$, ***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

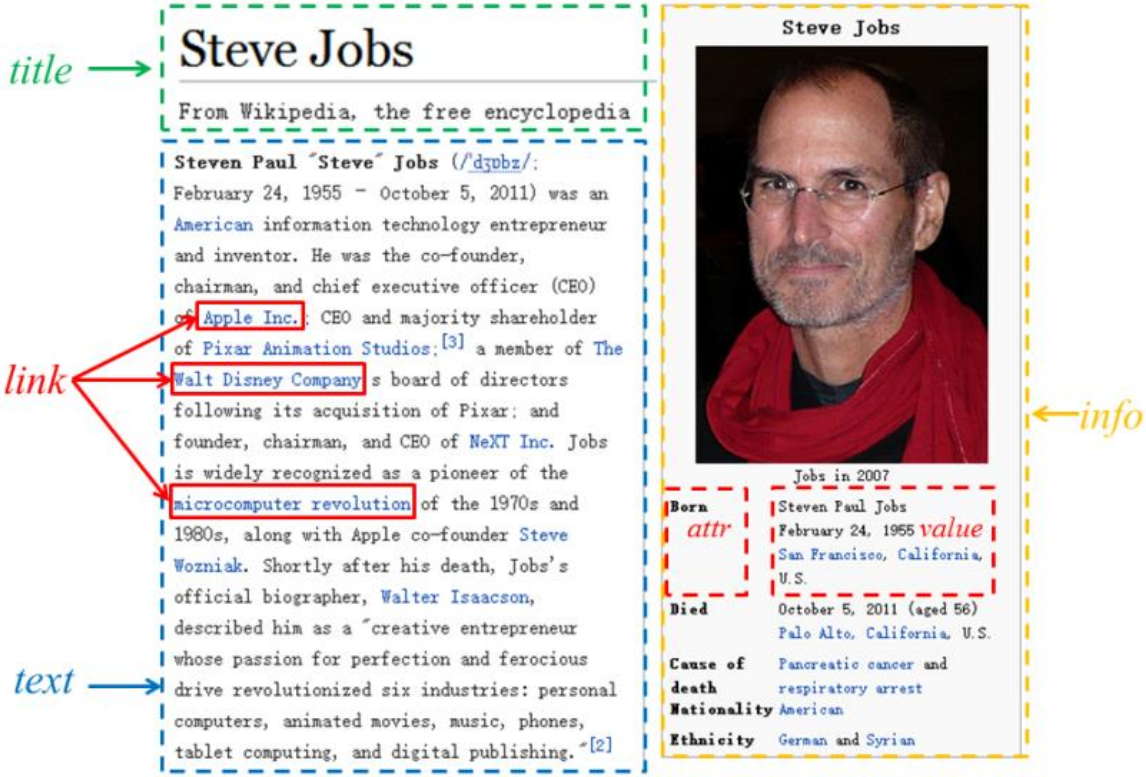
- Large number of known CLs are required in existed methods for cross-lingual knowledge linking for serving as either training data or seed set.
- There are less existing CLs or none at all between different wikis.

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***.



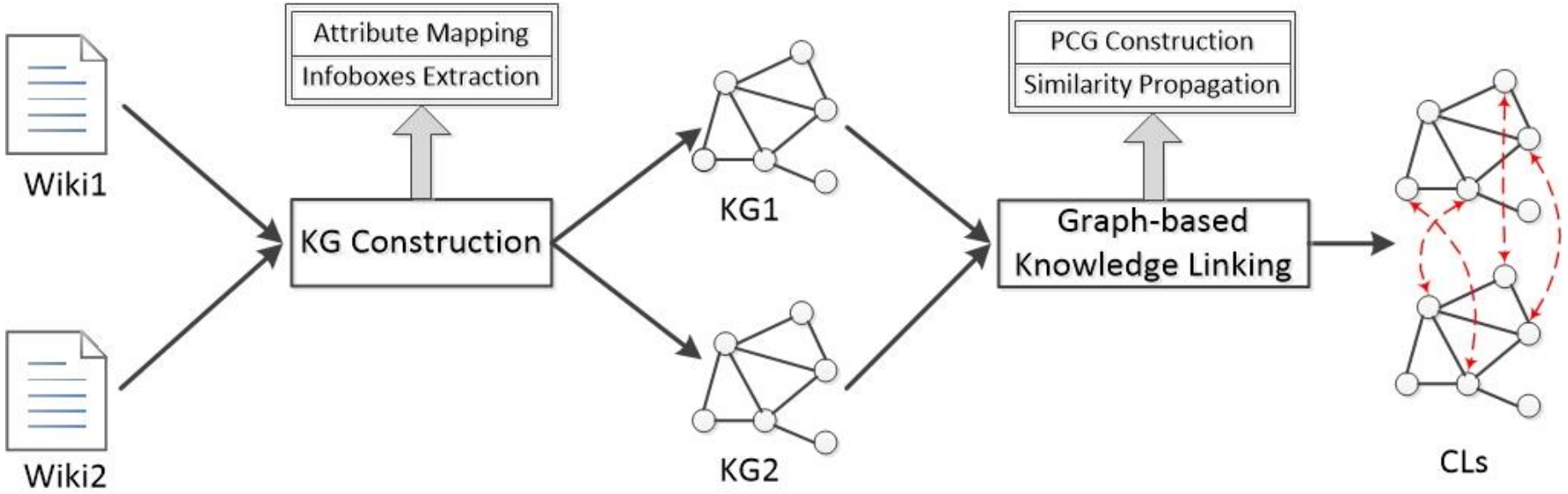
If we focus the problem in a given domain, ***Infoboxes*** can also be utilized for cross-lingual knowledge linking.

Outline

- Backgrounds
- Problem Definition
- **Methods**
- Experiments and Analysis
- Conclusion

The Proposed Approach

General Framework



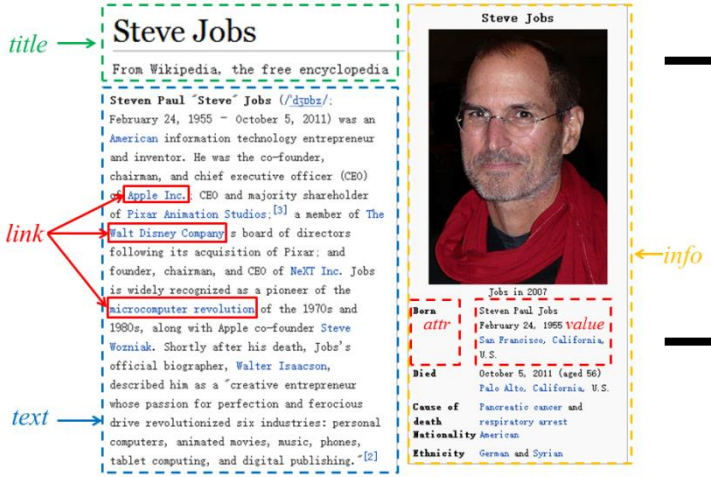
- Step1: **Knowledge Graph Construction**
  - Extract semantic relations in a structured form of subject-predicate-object triples from infoboxes of the two input wikis to construct two knowledge graphs.
- Step2: **Graph-based Knowledge Linking**
  - Discover CLs between the two constructed graphs based on a variation of the Similarity Flooding algorithm.

The Proposed Approach

Knowledge Graph Construction

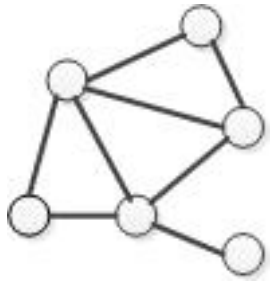
Input 1: Attribute Mapping of a given domain

Input 2: Wiki articles



Infoboxes

Inner Links

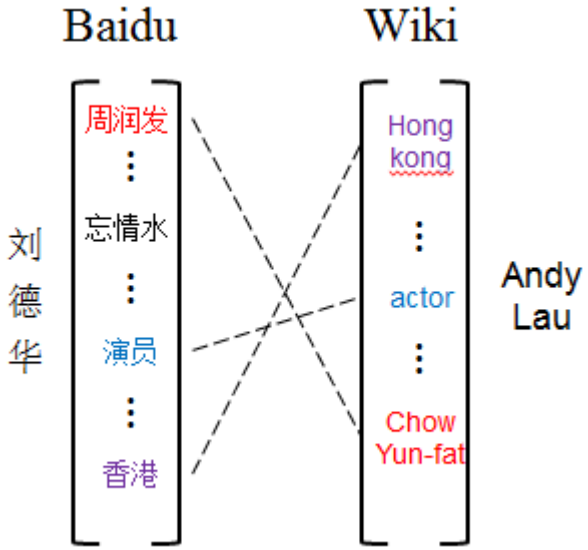


Knowledge Graph

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.

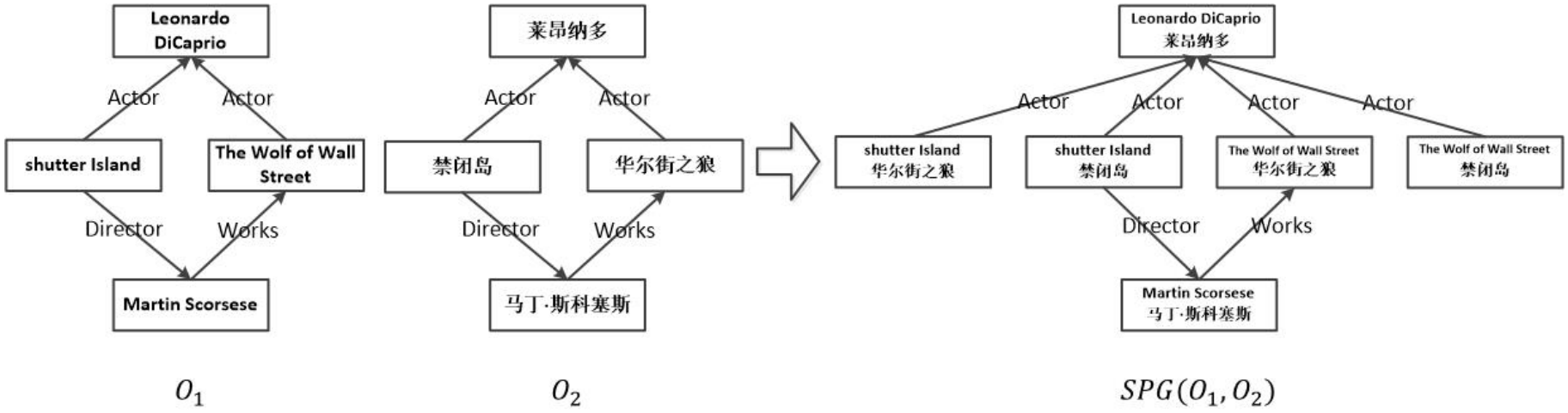


The Proposed Approach

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:

    (1) $(e, r, o) \in G$ (2) $(e', r, o') \in G'$ (3) $Sim(e, e') > \theta$ (4) $Sim(o, o') > \theta$
  - The $Sim(e, e')$ indicates the initial similarity between entity e and e'. θ is a pre-given threshold.



O_1 O_2 $SPG(O_1, O_2)$

The Proposed Approach

- Graph-based Knowledge Linking
  - Step 3: **Similarity Flooding**

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

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

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

- φ^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.
- ω(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