Error Link Detection and Correction in Wikipedia

Chengyu Wang, Rong Zhang, Xiaofeng He, Aoying Zhou

School of Computer Science and Software Engineering
East China Normal University
Shanghai, China

Outline

- **Introduction**
- Related Work
- Proposed Approach
- Experiments
- Conclusion

Introduction (1)

- Hyperlinks in Wikipedia
  - The hyperlink network in Wikipedia is valuable for knowledge harvesting, entity linking, etc.
  - Errors in the network structure are almost unavoidable and difficult to detect.
  - Goal of this paper: detect and correct error links in Wikipedia automatically.

Wikipedia	#Entities	#Links
English	3.6M	92M
Chinese	0.9M	11M

Java

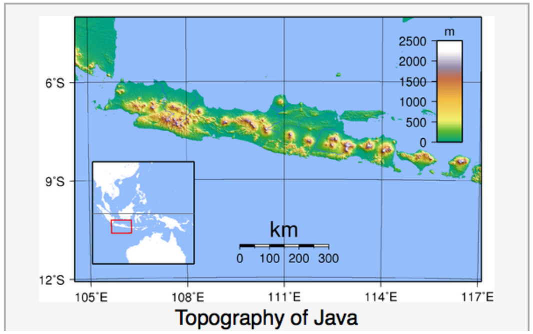
From Wikipedia, the free encyclopedia

Coordinates: 7°29′30″S 110°00′16″E

This article is about the Indonesian island. For the programming language, see Java ... other uses, see Java (disambiguation).

...se: □□;
...donesia. With
...he island
...ative region)
...ecember
...cent of the
...most
...lonesian
...western Java.
...lace on Java.
...u-Buddhist
...nd the core of

Java
Jawa (Indonesian)
□□ (Javanese)
□□ (Sundanese)



Topography of Java

Facebook

Links to

From Wikipedia, the free encyclopedia

Coordinates: 37.4848°N 122.1484°W

This article is about the social networking service. For the type of directory, see face book.

Facebook (stylized as **facebook**) is a for-profit corporation and online social media and social networking service based in Menlo Park, California, United States. The Facebook website was launched on February 4, 2004, by Mark Zuckerberg, along with fellow Harvard College students and roommates, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, and Chris Hughes.[7][8][9]

Facebook, Inc.

After this, data is output in PHP format (compiled with HipHop for PHP). The backend is written in Java and Thrift is used as the messaging format so PHP programs can query Java services. Caching solutions are used to make the web pages display more quickly. The more and longer data is cached the less realtime it is. The data is then sent to MapReduce servers so it can be queried via Hive. This also serves as a backup plan as the data can be recovered from Hive. Raw logs are removed after a period of time.[185]

The backend is written in Java…

...rogramming language)

...a, the free encyclopedia

...uage" redirects here. For the natural language from the Indonesian island of Java, ...ese language.

...e is about a programming language. For the software package downloaded from ... see Java SE.

...confused with JavaScript.

Correct!

Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented,[14] and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA),[15] meaning that compiled Java code can run on all platforms that support Java without the need for recompilation.[16] Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages in use,[17][18][19][20] particularly for client-server web applications, with a reported 9 million ...

Java

Java™

华东师范大学数据科学与工程研究院
Institute for Data Science and Engineering at ECNU

Introduction (2)

- Challenges
  - Error sparsity
    - A small number of error links v.s.10M+ Wikipedia links
  - Non-existent ground truth assumption
    - Wikipedia is treated as “ground truth” in traditional EL research.
    - No human-annotated error links are available.
- Two-stage Approach
  - Stage 1: generate candidate error links from Wikipedia with higher error density
  - Stage 2: predict error links and provide corrections at the same time

Outline

- Introduction
- **Related Work**
- Proposed Approach
- Experiments
- Conclusion

Related Work (1)

- Entity linking (EL)
  - Link an entity mention in text to a named entity in knowledge base
  - Methods: textual similarity, classification, learning to rank, graph-based ranking, etc.
  - Limitations
    - Wikipdia can not serve as the knowledge base for EL.
    - It is computationally costly to link all the anchor texts to Wikipedia pages.

Related Work (2)

- Wikification
  - Add links in documents to Wikipedia
  - A generalized task of EL
- Error link detection in Wikipedia
  - Pateman and Johnson's method
    - Highlight Wikipedia linking errors by analyzing the "semantic contribution" of Wikipedia links

Outline

- Introduction
- Related Work
- **Proposed Approach**
- Experiments
- Conclusion

General Framework

Two-stage Approach

- Candidate Error Link Generation
  - Construct a dictionary $M = \{(m, E_m)\}$ containing pairs of an anchor text m and its referent entity collection E_m
    - "Java": Java, Java (programming language)
  - Generate candidate error link set $CL_m = \{< l_{i,j}, l_{i,j'} >\}$ containing pairs of a candidate error link $l_{i,j}$ and its most possible correction $l_{i,j'}$
    - "Java": Facebook → Java, Facebook → Java (programming language)
- Link Classification and Correction
  - Train a classifier f to predict whether $l_{i,j}$ is an error link and $l_{i,j'}$ is a corrected link simultaneously
    - Error link: Facebook → Java
    - Corrected link: Facebook → Java (programming language)

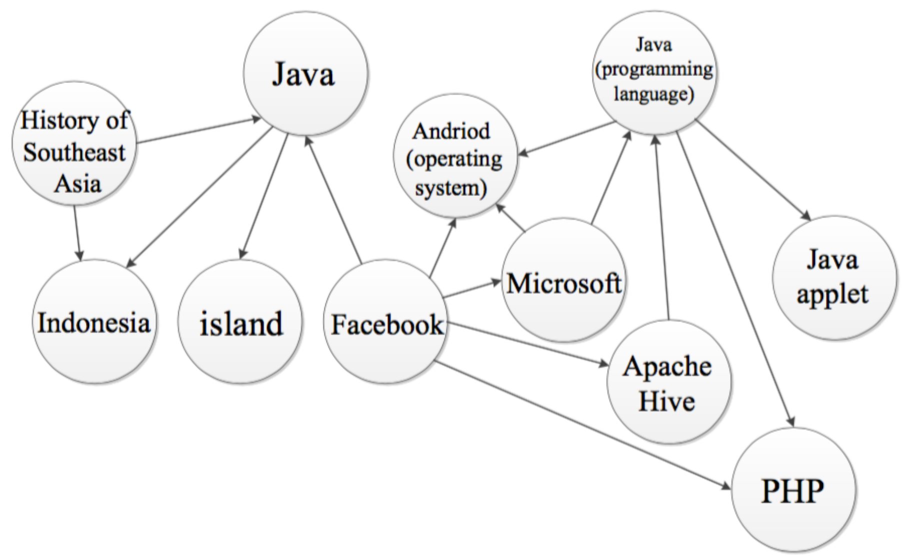
Candidate Error Link Generation

Dictionary and ATSN

- Dictionary Construction
  - Utilize Wikipedia to construct ambiguous anchor text-referent entity dictionary
    - Sources: redirect pages, disambiguation pages, hyperlinks, etc.
  - Example

Anchor Text m	Possible Referent Entity Collection E_m
Java	Java Java (programming language) ...
New York	New York City New York (magazine) New York (film) ...

- ATSN (Anchor Text Semantic Network)
  - For each anchor text
    - Nodes: referent entities and their neighbors
    - Links: hyperlinks between nodes



Candidate Error Link Generation

LinkRank Algorithm

- LinkRank
  - A PageRank-like algorithm to assign weights to links in an ATSN
  - Weight transition:
    - Links with non-zero outdegrees: pass weights to outlinks

$$u_{i,j}^{(n)} = \frac{1}{|OutLink_j|} \cdot w_{i,j}^{(n-1)}$$

    - Links with zero outdegree: distribute weights to all links uniformly
  - Weight update rule
    - Transitional weights + weights from zero out-degree links

$$w_{i,j}^{(n)} = \sum_{l_{k,i} \in InLink_i} u_{k,i}^{(n)} + \frac{1}{|L_m|} \sum_{l_{p,q} \in \bar{L}_m} w_{p,q}^{(n-1)}$$

Candidate Error Link Generation

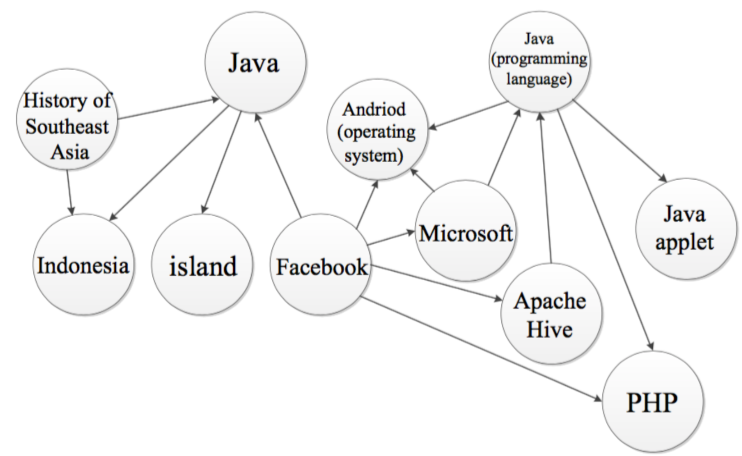
Set Generation

- Semantic Closeness (SC) between Two Entities in a Link
  - An asymmetric measurement based on LinkRank
  - SC from e_i to e_j: sum of weights of links between e_i and all e_j's neighbors

$$SC(e_i \rightarrow e_j) = \sum_{e_{j'} \in Neighbor(e_j) \wedge l_{i,j'} \in L_m} w_{i,j'}$$

- Criterion for candidate error link generation (three necessary conditions)
  - e_j and $e_{j'}$ share the same entity mention
  - e_i links to e_j in Wikipedia
  - Given a pre-defined threshold τ, we have

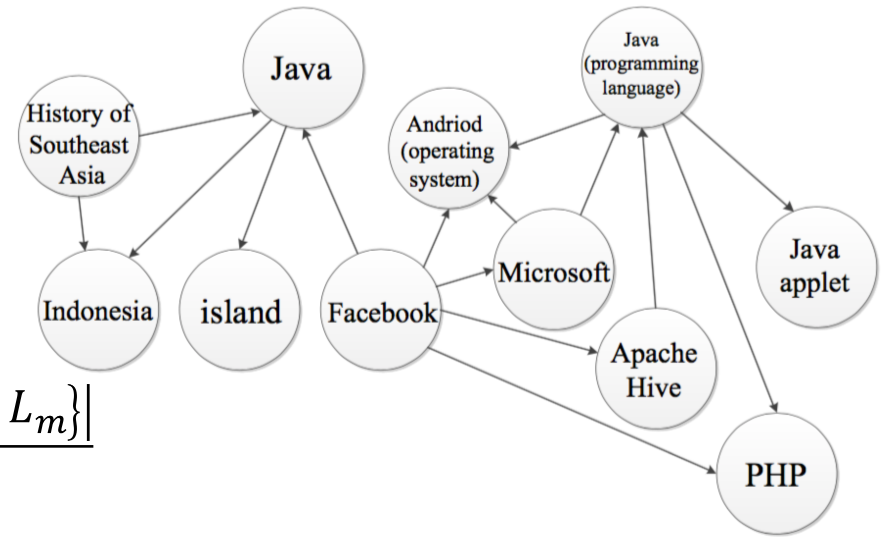
$$\frac{SC(e_i \rightarrow e_{j'}) - SC(e_i \rightarrow e_j)}{SC(e_i \rightarrow e_{j'})} > \tau$$



Link Classification and Correction

Feature Sets of a Link

- Graph-based Features
  - Inlink similarity
  - $ILS(i,j) = \frac{|InLinkNode_i \cap InLinkNode_j|+1}{|InLinkNode_i \cup InLinkNode_j|+1}$
  - Outlink similarity $OLS(i,j)$
  - Inlink relatedness
  - $ILR(i,j) = \frac{\left|\{e_k \in InLinkNode_i | l_{k,j} \in L_m\}\right|}{|InLinkNode_i|}$
  - Outlink relatedness $OLR(i,j)$



- Context-based Features
  - Context similarity $CS(i,j) = \frac{s_i^T \cdot s_j}{\|s_i\|_2 \cdot \|s_j\|_2}$
  - Frequent context similarity $FCS(i,j) = \frac{FS_i^T \cdot FS_j}{\|FS_i\|_2 \cdot \|FS_j\|_2}$

Link Classification and Correction

Pairwise Learning

- Feature Vector Construction
  - Feature vector of a link $l_{i,j}$

    $$v(l_{i,j}) =< ILS(i,j), OLS(i,j), ILR(i,j), OLR(i,j), CS(i,j), FCS(i,j) >$$
  - Vector difference between two links: $v_S(l_{i,j}, l_{i,j'}) = v(l_{i,j}) - v(l_{i,j'})$
  - Feature vector of a data instance: $v_{PL}(l_{i,j}, l_{i,j'}) =< v(l_{i,j}), v(l_{i,j'}), v_S(l_{i,j}, l_{i,j'}) >$
  - Example
    - Facebook → Java: 6 features
    - Facebook → Java (programming language): 6 features
    - The data instance: 6+6+6=18 features
- Pairwise Learning
  - Train a SVM classifier f to predict whether $l_{i,j}$ is an error link and $l_{i,j'}$ is a corrected link based on $v_{PL}(l_{i,j}, l_{i,j'})$

Outline

- Introduction
- Related Work
- Proposed Approach
- **Experiments**
- Conclusion

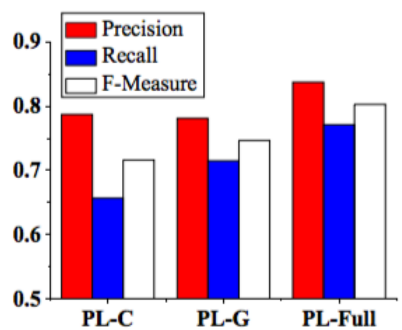
Experiments (1)

- Datasets: English and Chinese Wikipedia dumps
- Candidate Error Link Generation
  - Sample candidate error links and compare the density of error links
  - Methods for comparison
    - **Simple**: extract links that connects ambiguous entities based on disambiguation pages
    - **AnchorText**: extract links with ambiguous anchor texts based on the dictionary
    - **Unweighted**: the proposed approach with uniform link weights
    - **LinkRank**: the proposed approach with varied parameter settings

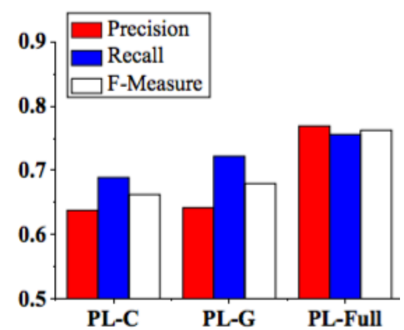
Method	# Error links in sample set	Density of error links
Dataset: English Wikipedia		
Simple	0	0% (approx.)
AnchorText	0	0% (approx.)
Unweighted	21	4.2%
LinkRank ($\tau = 0.2$)	28	5.6%
LinkRank ($\tau = 0.4$)	34	6.8%
LinkRank ($\tau = 0.6$)	43	8.6%
LinkRank ($\tau = 0.8$)	**58**	**11.6%**
Dataset: Chinese Wikipedia		
Simple	0	0% (approx.)
AnchorText	1	0.2%
Unweighted	17	3.4%
LinkRank ($\tau = 0.2$)	20	4.0%
LinkRank ($\tau = 0.4$)	26	5.2%
LinkRank ($\tau = 0.6$)	38	7.6%
LinkRank ($\tau = 0.8$)	**42**	**8.4%**

Experiments (2)

- Link Classification and Correction
  - Use SVM as the classifier to train models on candidate error link sets
  - Methods for comparison (considering feature subsets)
    - PL-C: use context-based features only
    - PL-G: use graph-based features only
    - PL-Full: use both context-based and graph-based features



English Wikipedia



Chinese Wikipedia

Experiments (3)

- Comparison between PL-Full and other methods

1. VSM: Compare content similarity based on Vector Space Model
2. EL: Link ambiguous anchor texts to referent entities in Wikipedia
3. LS: Detect incorrect links based on Wikipedia link structure
4. ELD: Use a classifier to predict error links directly (w/o pairwise learning)

Category	Method	Precision	Recall	F-Measure
Dataset: English Wikipedia				
VSM based	**VSim**	53.2%	40.8%	46.2%
	IntroVSim	57.9%	53.2%	55.5%
EL based	**Wikify!** [14]	45.4%	48.9%	47.1%
	LINDEN [24]	46.5%	61.4%	52.9%
Error link detection based	**LS** [17]	71.4%	58.6%	64.4%
	ELD	76.9%	47.3%	58.6%
	PL-Full	**83.7%**	**77.1%**	**80.3%**
Dataset: Chinese Wikipedia				
VSM based	**VSim**	50.1%	42.1%	45.8%
	IntroVSim	56.3%	51.2%	53.6%
EL based	**Wikify!** [14]	48.2%	41.5%	44.6%
	LINDEN [24]	43.8%	38.6%	41.0%
Error link detection based	**LS** [17]	68.5%	62.3%	65.3%
	ELD	54.7%	39.7%	46.0%
	PL-Full	**76.9%**	**75.6%**	**76.2%**

Analysis of Error Links

- Different types of ambiguity
  - MSNE: Multiple Senses of Named Entities
    - Error link: Josh White → Bob Gibson
    - Correction: Bob Gibson (musician)
  - MSC: Multiple Senses of Concepts
    - Error link: Cheltenham Town F.C. → Administration (law)
    - Correction: Administration (British football)
  - ACNE: Ambiguity Between Concepts and Named Entities
    - Error link: Tactical role-playing game → Steam
    - Correction: Steam (software)

Dataset	Category of error links		
	MSNE	ACNE	MSC
Wikipedia Error Link Set (English)	**75.8%**	20.8%	3.4%
Wikipedia Error Link Set (Chinese)	**83.6%**	11.8%	4.6%

Case Studies

- English Wikipedia

Category	Source Wikipage	Target Wikipage	Correct Wikipage
MSNE	Augustus of Prima Porta[1]	Mars	Mars (mythology)
	Josh White	Bob Gibson	Bob Gibson (musician)
MSC	Cheltenham Town F.C.	Administration (law)	Administration (British football)
ACNE	Tactical role-playing game	Steam	Steam (software)
	Ireland in the Eurovision Song Contest 2011[2]	Lipstick	Lipstick (Jedward song)

- Chinese Wikipedia

Category	Source Wikipage	Target Wikipage	Correct Wikipage
MSNE	Theodore Beza[1] (泰奥多尔·贝扎)	Baden (巴登)	Baden (Switzerland) (巴登 (瑞士))
	Light Rail 705 & 706[2] (香港轻铁705、706线)	Ginza Station (银座站)	Ginza Stop (Hong Kong) (银座站 (香港))
MSC	Unit sphere[3] (单位球面)	Boundary (边界)	Boundary (topology) (边界 (拓扑学))
ACNE	Donnie Yen[4] (甄子丹)	Hero (英雄)	Hero (film) (英雄 (电影))
	Zhou Yang (actress)[5] (周扬 (演员))	Tea house (茶馆)	Tea House (TV series) (茶馆 (电视剧))

Outline

- Introduction
- Related Work
- Proposed Approach
- Experiments
- **Conclusion**

Conclusion

- Methods
  - The two-stage approach is effective to detect and correct error links in Wikipedia.
    - Stage 1: generate candidate error links with higher density
    - Stage 2: predict error links and provide corrections at the same time
- Analysis
  - Most linking errors in Wikipedia are caused by multiple senses of named entities.
- Future work
  - Detecting error links where the correct entities is outside Wikipedia.
  - Detecting and correcting errors in other Web-scale networks.

Thanks!

Questions & Answers

* The first author would like to thank CIKM 2016 for the SIGIR student travel grant.