Building Chinese Linking Open Data

FANG Yishu (246*****)
XIA Yingying (246*****)

SEU-MONASH Joint Graduate School
31 May 2013

Content

- ***Introduction*** ***– LOD & CLOD***
- ***Objective***
- ***Methodology*** ***– Data acquiring, representation, and processing***
- ***Novelty & Conclusion***
- ***References***

Introduction - LOD

- ***What is Linked Data?***
- ***LOD – Linking Open Data***

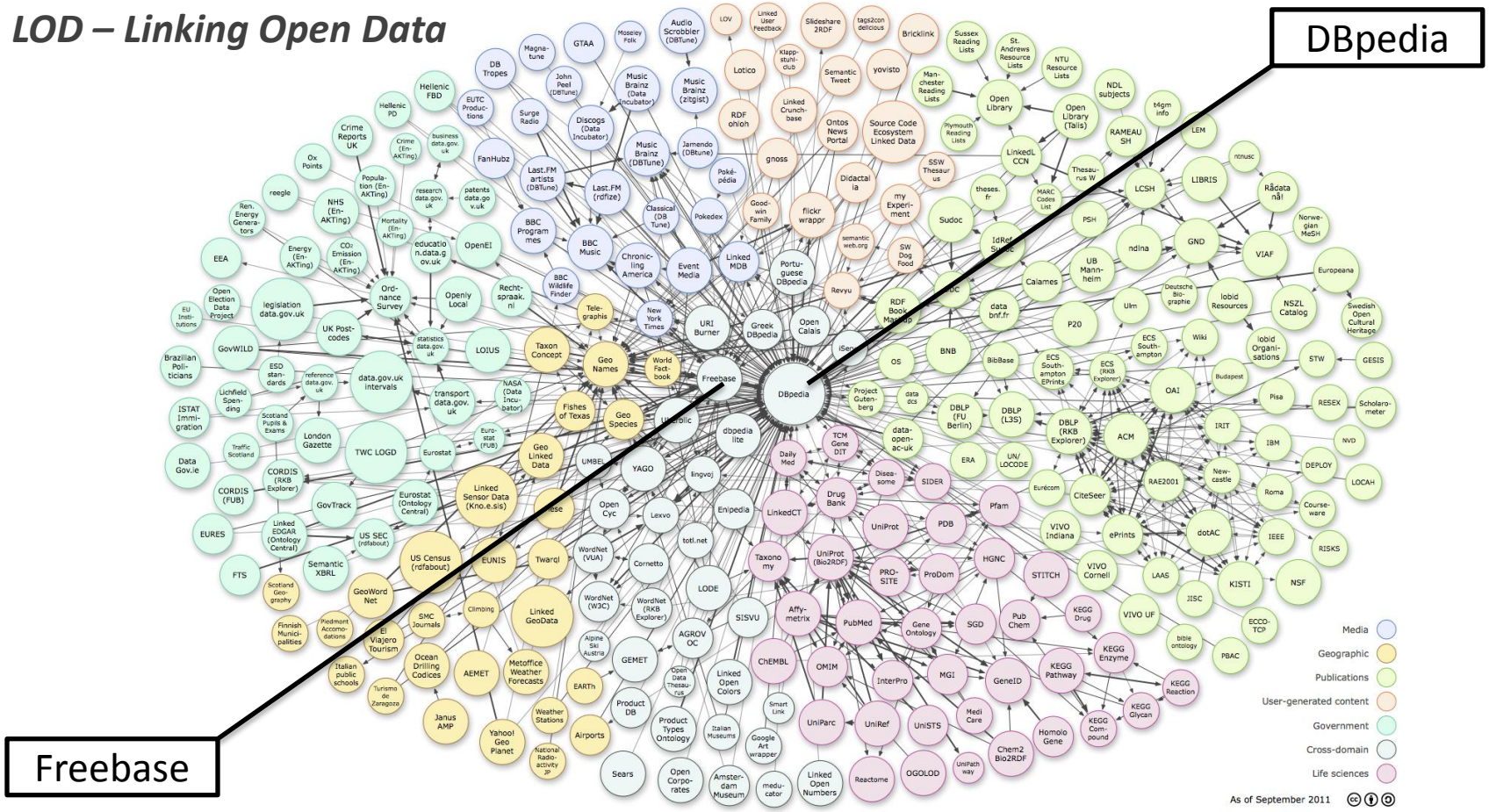


Figure: LOD Cloud Diagram as of September 2011 (Each node is a dataset)

Introduction - CLOD

- *CLOD – Chinese Linking Open Data*

- ***Problem 1: native CLOD dataset rarely exists***
  - *Only two datasets exist*
- ***Problem 2: Chinese information also rarely appears in existing multilingual LOD datasets***
  - *They are built on English with some entries translated to Chinese*
  - *Not complete and lack of real-world usage*

Objectives

Build a LOD dataset native to Chinese.

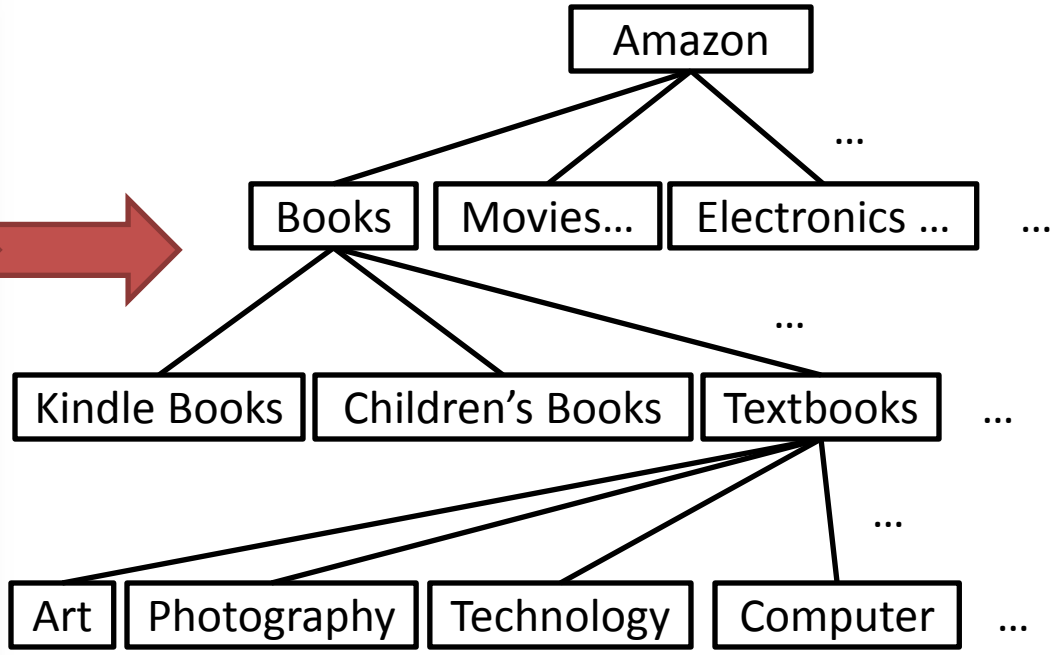
Two representative problems:

1. ***How to discover synonyms***
2. ***How to merge ontologies***

Methodology

1. *Category Extraction*
2. *Category Representation*
3. *Synonym Discovery*
4. *Ontology Merging*

Methodology – Category Extraction



We have extracted

Category hierarchy from **50** Chinese sites, then **50** trees (**ontologies**) have been built.

e.g.: 淘宝网，亚马逊（中国），维基百科（中文），互动百科……

Methodology – Category Representation

Three ways to represent a category:

1. *String of the name*
2. *Related category*
3. *Related category frequency*

e.g.:

1. 南京
2. 南京：江苏 **地区** 中国历史文化名城 **香烟品牌** 石头城 古都 …
3. 南京：江苏 8 地区 8 中国历史文化名城 1 香烟品牌 1 石头城 1 古都 1 …

Methodology – Synonym Discovery

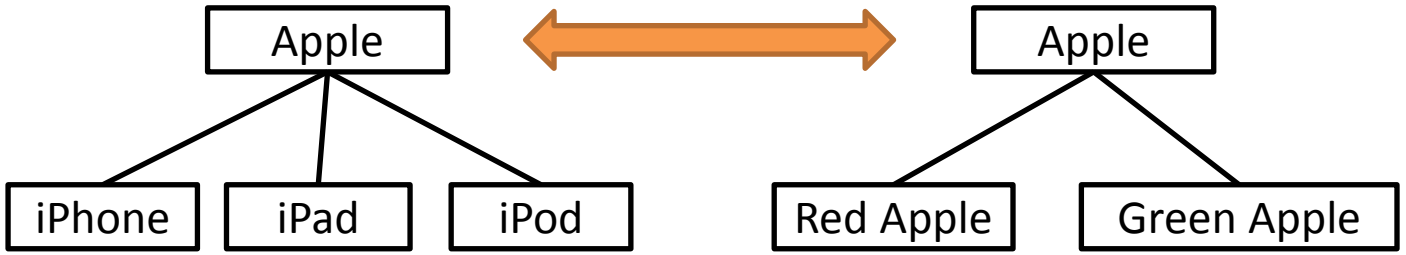
Three algorithms to estimate how two words are similar to each other

1. ***BOWISim:*** *based on Levenshtein distance*
2. ***BORCSSim:*** *based on Jaccard similarity*
   (work on Related category)
3. ***BORCVSim:*** *based on cosine similarity*
   (work on Related category frequency)

e.g.: *New York City vs. NYC*

Methodology – Ontology Merging

- ***To determine the `owl:sameAs` relationship, synonym is not enough to combine two nodes together.***
- e.g.:



- ***So we also need to inspect their* context information**

Novelty & Conclusion

We have built a LOD dataset native to Chinese

Two problems have been solved:

1. ***How to discover synonyms***
2. ***How to merge ontologies***

References

- Berners-Lee, T., & O'Hara, K. (2013). The read–write Linked Data Web. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1471-2962.
- Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., & Yu, Y. (2011). *Zhishi. me-weaving Chinese linking open data.* Paper presented at the The Semantic Web–ISWC 2011, 205-220.
- Schimke, S., Vielhauer, C., & Dittmann, J. (2004). *Using adapted levenshtein distance for on-line signature authentication.* Paper presented at the Proceedings of the 17th International Conference on Pattern Recognition, 931-934.
- Shen, D., Qin, M., Chen, W., Yang, Q., & Chen, Z. (2007). *Mining web query hierarchies from clickthrough data.* Paper presented at the Proceedings of the National Conference on Artificial Intelligence, 341-346
- Strehl, A., Ghosh, J., & Mooney, R. (2000). *Impact of similarity measures on web-page clustering.* Paper presented at the Workshop on Artificial Intelligence for Web Search (AAAI 2000), 58-64.
- Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1091-1095.
- Zhao, J. (2010). Publishing Chinese medicine knowledge as Linked Data on the Web. *Chinese medicine, 5*(1), 1-12.