東南大學

毕业设计（论文）任务书

院（系）　计算机科学与工程学院

专　　业　　计算机科学与技术

设计（论文）题目　**基于多知识库的表格实体链接系统**

学 生 姓 名　严晟嘉　学号　09013119

起 止 日 期　2017 年 1 月 ～ 2017 年 6 月

设 计 地 点　东南大学九龙湖校区

院 内 导 师　李慧颖

院 外 导 师

教学院长（教学系主任）

发任务书日期　　年　　月　　日

毕业设计（论文）任务的内容和要求

（包括任务内容、原始条件及数据、技术要求、工作要求等）

任务内容

WWW(世界万维网)中包含了超大规模的有价值的关系数据，它们很多都存在于 HTML 表格中。为了从这些表格中抽取出有价值的信息，关键的一步就是实体链接技术。如此大规模的数据已经达到了过载的程度，这就导致了人们在找到有用的信息之前，我们需要处理更多的无用信息并且找到多少有用的信息取决于我们从沙子里淘金的本领。人们不能高效地找到有用的信息的原因就在于自然语言表达的多样性和歧义性。我们的任务就是实现一个基于多知识库的表格实体链接系统来解决上述问题。

原始条件及数据

原始条件：

* 该课题旨在提出一种统一的架构以解决不同类型数据的中的实体链接任务，并且期望效果要优于或至少持平于目前最先进的实体链接技术。
* 现有的工作都是基于单一知识库进行实体链接，而实际上任何知识库都是不完整的，我们的工具期望支持多种知识库进行实体链接，从而缓解知识库的完整性问题。

数据：

数据分为中文数据和英文数据。

中文数据来自最大的中文链接知识库 Zhishi.me，其包含了三个中文百科知识库：中文维基百科、百度百科和互动百科。

英文数据来自维基百科。

技术要求

* 网络爬虫技术。研究的前期准备就是爬取社交网络（微博，微信公众号等）中的数据。
* 数据预处理技术。预处理包括对知识库优化、同义词表构建、百科词条访问量提取和微博语句预处理。
* 表格数据实体链接技术。首先查找知识库判断是否查找成功, 若没有则 找同义词表, 若同义词表中不存在则采用改进的拼音编辑距离算法（或者其他距离算法）进行知识库查找, 若仍没有查找到则采用后缀词表匹配法（或者其他匹配法）查找知识库。具体使用的算法要依实际情况决定。
* 实体消歧技术。使用一些聚类算法或者 SVM 支持向量机模型来消除歧义。

工作要求

1、在深刻领会任务内容及要求的基础上，通过查阅文献资料、调查研究和方案论证，写出开题报告。然后开展实验研究、理论研究、设计、研制、开发以及数据处理、分析总结、资

料整理等与任务书要求相应的工作，并撰写成毕业论文或设计说明书，独立地完成毕业设计的各项任务；

2、查找有关专业文献（10 篇以上）；

3、毕业论文或设计说明书需符合规范化要求，即：由中外文题名、目录、中外文摘要、引言（前言）、正文、结论、谢辞、参考文献和附录组成，中文摘要在 400 汉字左右，外文摘要在 250 个实词左右，中文题名字数一般不超过 20 个，设计说明书、论文或软件说明书的总字数在 1.5~2 万汉字（文、管等学科可根据具体情况，另行规定总字数，报教务处备案）。

学生应提交的软硬件的名称、内容及主要的技术指标:

计算机软件：一个基于多知识库的表格实体链接系统

应提交的其它文档：

（1）开题报告一份

（2）与设计（论文）相关的英文资料译文一份（中文字数>5000 字，并附保留阅读痕迹的资料原文）

（3）毕业设计论文一份

参考文献（至少五篇，含供学生翻译的英文资料，按规范开列）：

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: ISWC, pp. 722–735 (2007)

2. Bhagavatula, C.S., Noraset, T., Downey, D.: Tabel: Entity linking in web tables. In: ISWC, pp. 425–441 (2015)

3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hell- mann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: science, services and agents on the world wide web 7(3), 154–165 (2009)

4. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collabo- ratively created graph database for structuring human knowledge. In: SIGMOD, pp. 1247–1250 (2008)

5. Brin,S.,Page,L.:Reprintof:Theanatomyofalarge-scalehypertextualwebsearch engine. Computer networks 56(18), 3825–3833 (2012)

6. Cafarella, M.J., Halevy, A., Wang, D.Z., Wu, E., Zhang, Y.: Webtables: exploring the power of tables on the web. PVLDB 1(1), 538–549 (2008)

7. Craswell, N.: Mean reciprocal rank. In: Encyclopedia of Database Systems, pp. 1703–1703. Springer (2009)

8. Ferna´ndez-Delgado,M.,Cernadas,E.,Barro,S.,Amorim,D.:Doweneedhundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15(1), 3133–3181 (2014)

9. Hignette, G., Buche, P., Dibie-Barth´elemy, J., Haemmerl´e, O.: Fuzzy annotation of web data tables driven by a domain ontology. In: ESWC, pp. 638–653 (2009)

10. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities,

types and relationships. PVLDB 3(1-2), 1338–1347 (2010)
11. Mulwad, V., Finin, T., Joshi, A.: Semantic message passing for generating linked data from tables. In: ISWC, pp. 363–378 (2013)
12. Mun˜oz, E., Hogan, A., Mileo, A.: Using linked data to mine rdf from wikipedia's tables. In: WSDM, pp. 533–542 (2014)
13. Navigli, R., Ponzetto, S.P.: Babelnet: Building a very large multilingual semantic network. In: ACL, pp. 216–225 (2010)
14. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi. me-weaving chinese linking open data. In: ISWC, pp. 205–220 (2011)
15. Pereira, B.: Entity linking with multiple knowledge bases: An ontology modular- ization approach. In: ISWC, pp. 513–520 (2014)
16. Shen, W., Wang, J., Luo, P., Wang, M.: Liege:: link entities in web lists with knowledge base. In: SIGKDD, pp. 1424–1432 (2012)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW, pp. 697–706 (2007)
18. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (2008)
19. Syed, Z., Finin, T., Mulwad, V., Joshi, A.: Exploiting a web of semantic data for interpreting tables. In: WebSci, vol. 5 (2010)
20. Venetis, P., Halevy, A., Madhavan, J., Pa¸sca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. PVLDB 4(9), 528–538 (2011)
21. Zhang, Z.: Learning with partial data for semantic table interpretation. In: EKAW, pp. 607–618 (2014)
22. Zhang, Z.: Towards e cient and e↵ective semantic table interpretation. In: ISWC, pp. 487–502 (2014)

毕业设计（论文）进度安排

起止日期	工作内容	备　注
2017-02-01 ~ 2017-02-28	阅读文献资料，思考算法模型	
2017-03-01 ~ 2017-03-31	准备数据集并做实验	

2017-04-01 ~ 2017-04-30	写论文做系统	
2017-05-01 ~ 2017-05-20	完善论文与系统	

注：只需按阶段作出安排，更细的安排应由学生自己在开题报告中作出。

院内导师签名：　　院外导师签名：

年　月　日