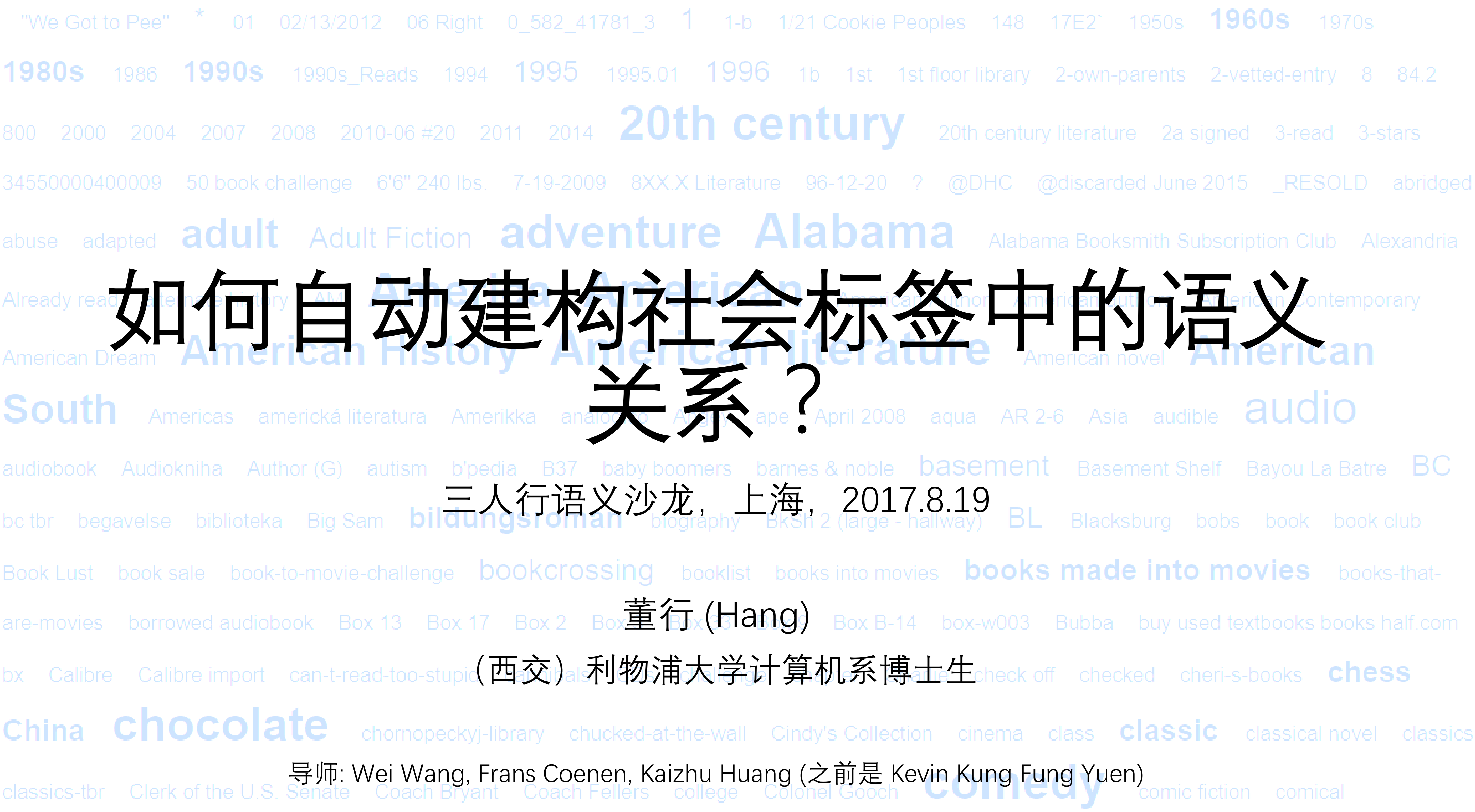

"We Got to Pee" * 01 02/13/2012 06 Right 0_582_41781_3 1 1-b 1/21 Cookie Peoples 148 17E2` 1950s 1960s 1970s
1980s 1986 1990s 1990s_Reads 1994 1995 1995.01 1996 1b 1st 1st floor library 2-own-parents 2-vetted-entry 8 84.2
800 2000 2004 2007 2008 2010-06 #20 2011 2014 20th century 20th century literature 2a signed 3-read 3-stars
34550000400009 50 book challenge 6'6" 240 lbs. 7-19-2009 8XX.X Literature 96-12-20 ? @DHC @discarded June 2015 _RESOLD abridged
abuse adapted adult Adult Fiction adventure Alabama Alabama Booksmith Subscription Club Alexandria
Already read American contemporary
American Dream American History American literature American novel American
South Americas americká literatura Amerikka April 2008 aqua AR 2-6 Asia audible audio
audiobook Audiokniha Author (G) autism b'pedia B37 baby boomers barnes & noble basement Basement Shelf Bayou La Batre BC
bc tbr begavelse biblioteka Big Sam bildungsroman biography BkSh 2 (large - hallway) BL Blacksburg bobs book book club
Book Lust book sale book-to-movie-challenge bookcrossing booklist books into movies books made into movies books-that-
are-movies borrowed audiobook Box 13 Box 17 Box 2 Box B-14 box-w003 Bubba buy used textbooks books half.com
bx Calibre Calibre import can-t-read-too-stupid check off checked cheri-s-books chess
China chocolate chornopeckyj-library chucked-at-the-wall Cindy's Collection cinema class classic classical novel classics
classics-tbr Clerk of the U.S. Senate Coach Bryant Coach Fellers college Colonel Gooch comedy comic fiction comical

如何自动建构社会标签中的语义关系？

三人行语义沙龙，上海，2017.8.19

董行 (Hang)

（西交）利物浦大学计算机系博士生

导师: Wei Wang, Frans Coenen, Kaizhu Huang (之前是 Kevin Kung Fung Yuen)

UNIVERSITY OF LIVERPOOL

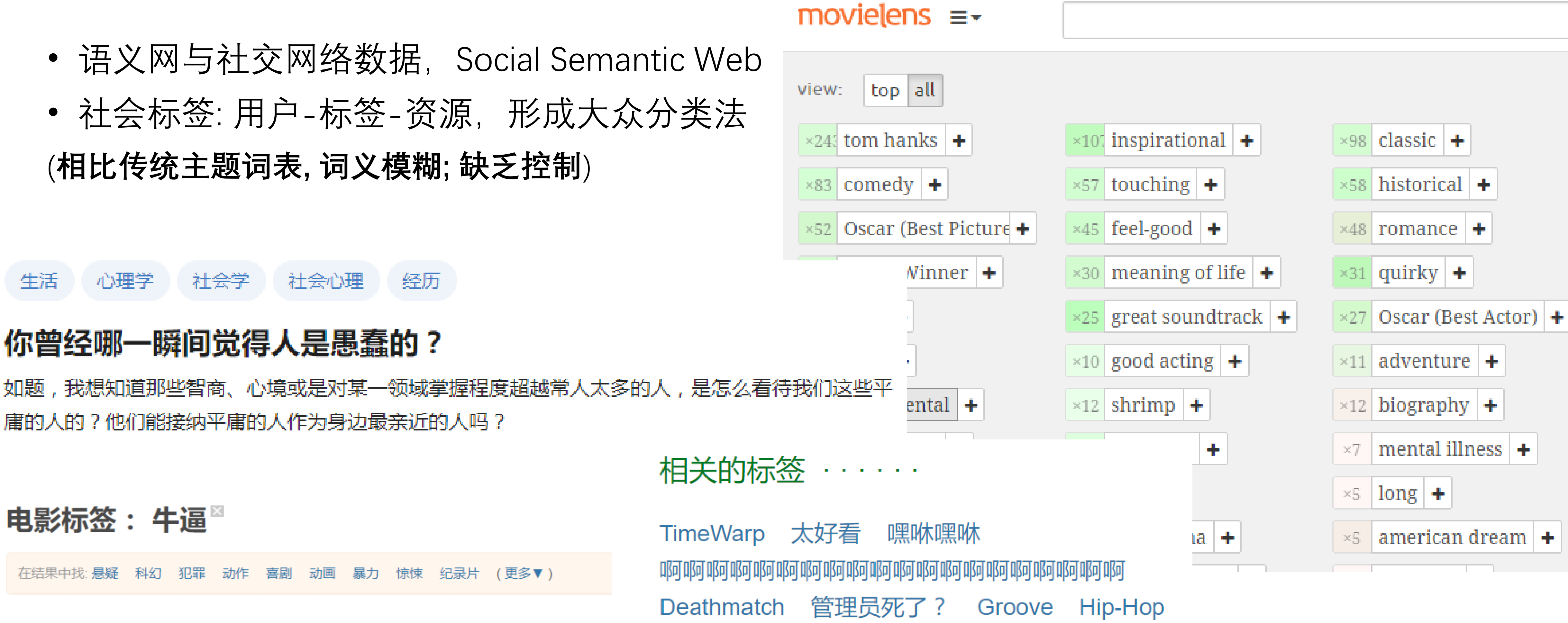
Xi’an Jiaotong-Liverpool University
西交利物浦大学

鸣谢本报告中用到的
各类网络与学术资源

从社交媒体数据中提取语义关系

- 语义网与社交网络数据，Social Semantic Web
- 社会标签: 用户-标签-资源，形成大众分类法

(**相比传统主题词表, 词义模糊; 缺乏控制**)



movielens

view: top all

×24 tom hanks +

×83 comedy +

×52 Oscar (Best Picture +

Winner +

ental +

×10 inspirational +

×57 touching +

×45 feel-good +

×30 meaning of life +

×25 great soundtrack +

×10 good acting +

×12 shrimp +

×98 classic +

×58 historical +

×48 romance +

×31 quirky +

×27 Oscar (Best Actor) +

×11 adventure +

×12 biography +

×7 mental illness +

×5 long +

×5 american dream +

生活 心理学 社会学 社会心理 经历

你曾经哪一瞬间觉得人是愚蠢的？

如题，我想知道那些智商、心境或是对某一领域掌握程度超越常人太多的人，是怎么看待我们这些平庸的人的？他们能接纳平庸的人作为身边最亲近的人吗？

电影标签： 牛逼

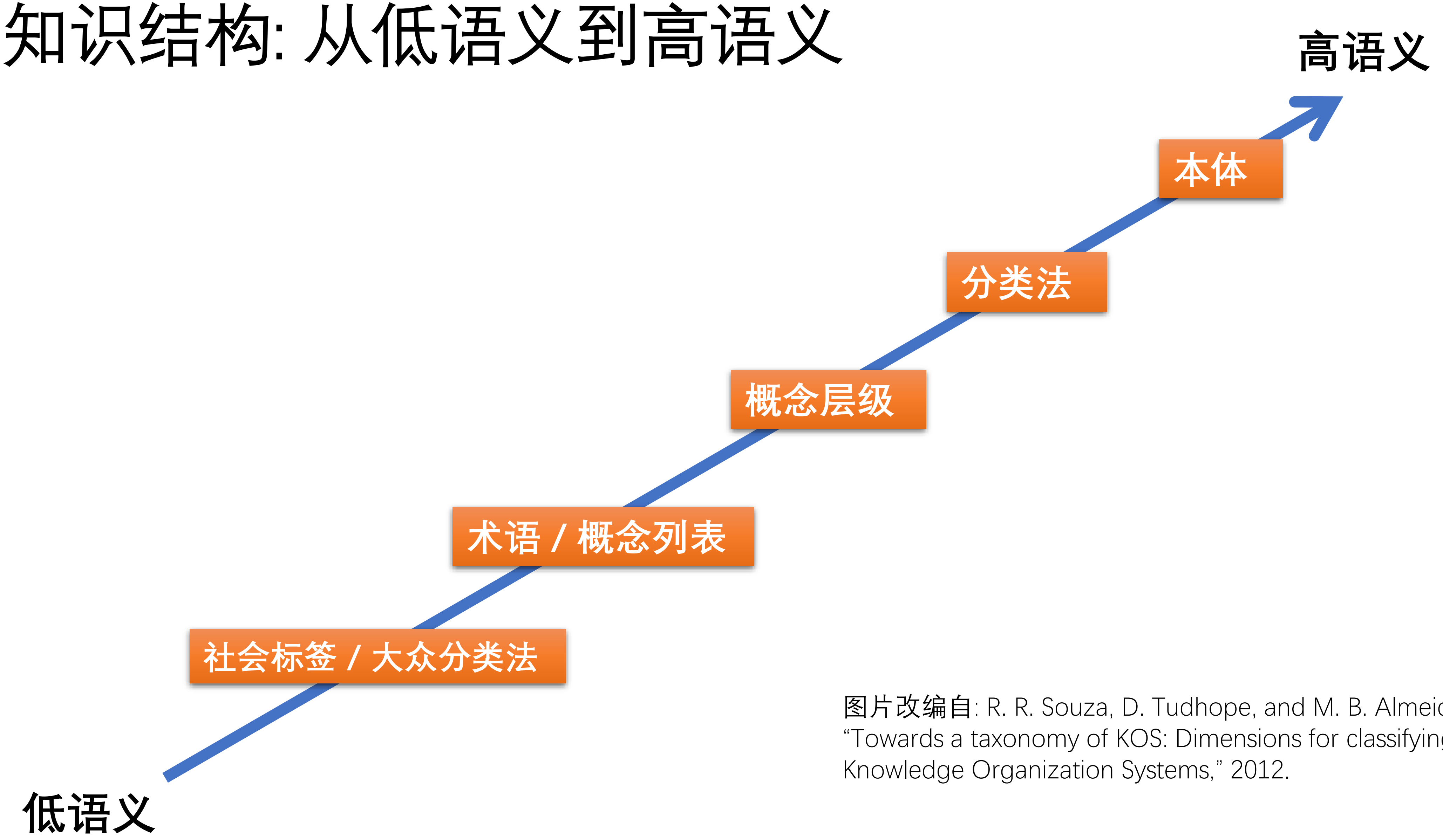
在结果中找: 悬疑 科幻 犯罪 动作 喜剧 动画 暴力 惊悚 纪录片 （更多▼）

相关的标签 · · · · · ·

TimeWarp 太好看 嘿咻嘿咻 啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊 Deathmatch 管理员死了？ Groove Hip-Hop

截图自: [1] https://www.zhihu.com/question/63444484/answer/216192705; [2] https://movielens.org/movies/356; [3] https://movie.douban.com/tag/%E7%89%9B%E9%80%BC

知识结构: 从低语义到高语义



图片改编自: R. R. Souza, D. Tudhope, and M. B. Almeida, “Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems,” 2012.

本体学习 Ontology learning

- 建立类似分类法的知识结构需要大量的人力和时间

- **从自然语言文本中自动化或者半自动化地建立本体**

- 社交网络中产生的新语言往往不被现有的分类体系收入，为本体学习提供了新的需求和素材

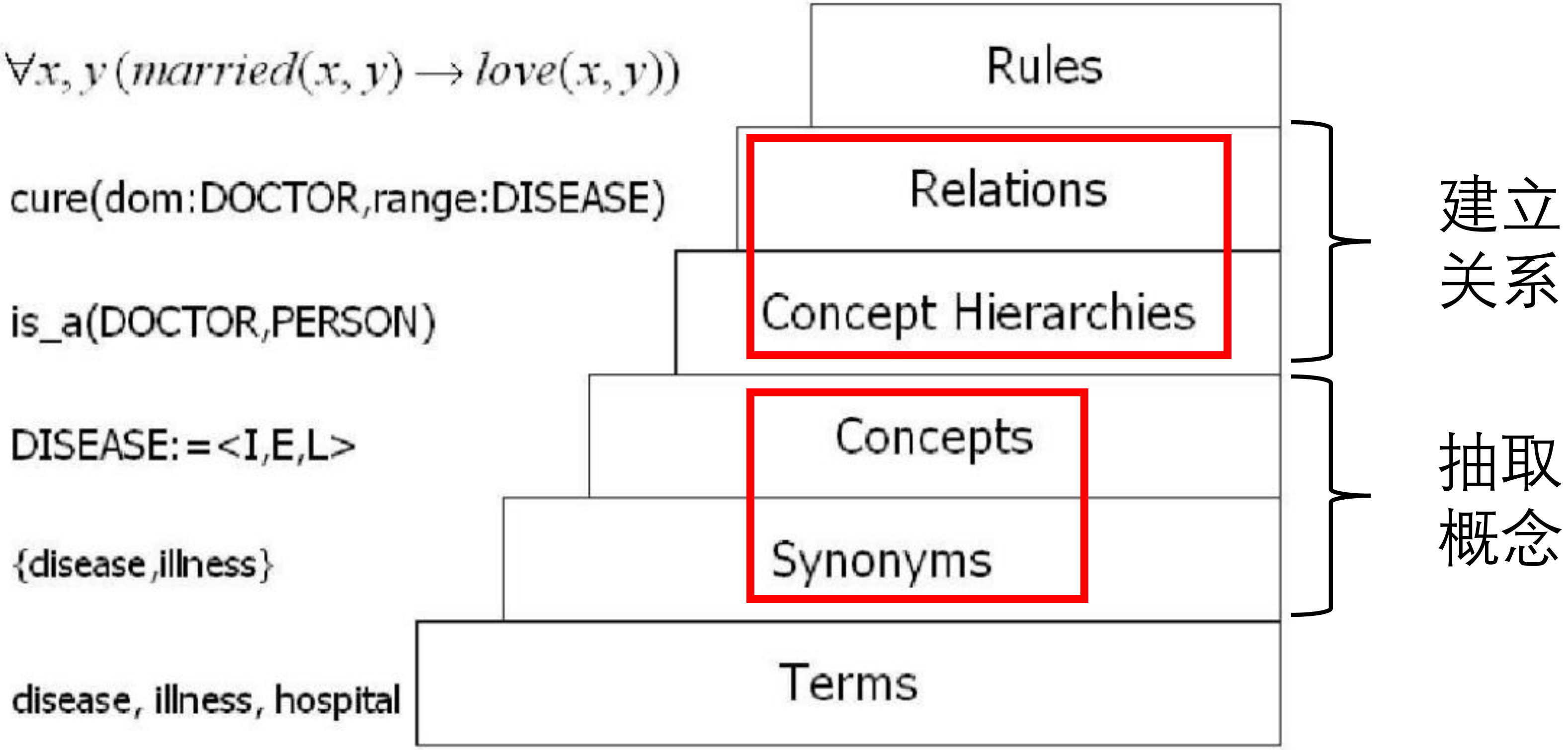
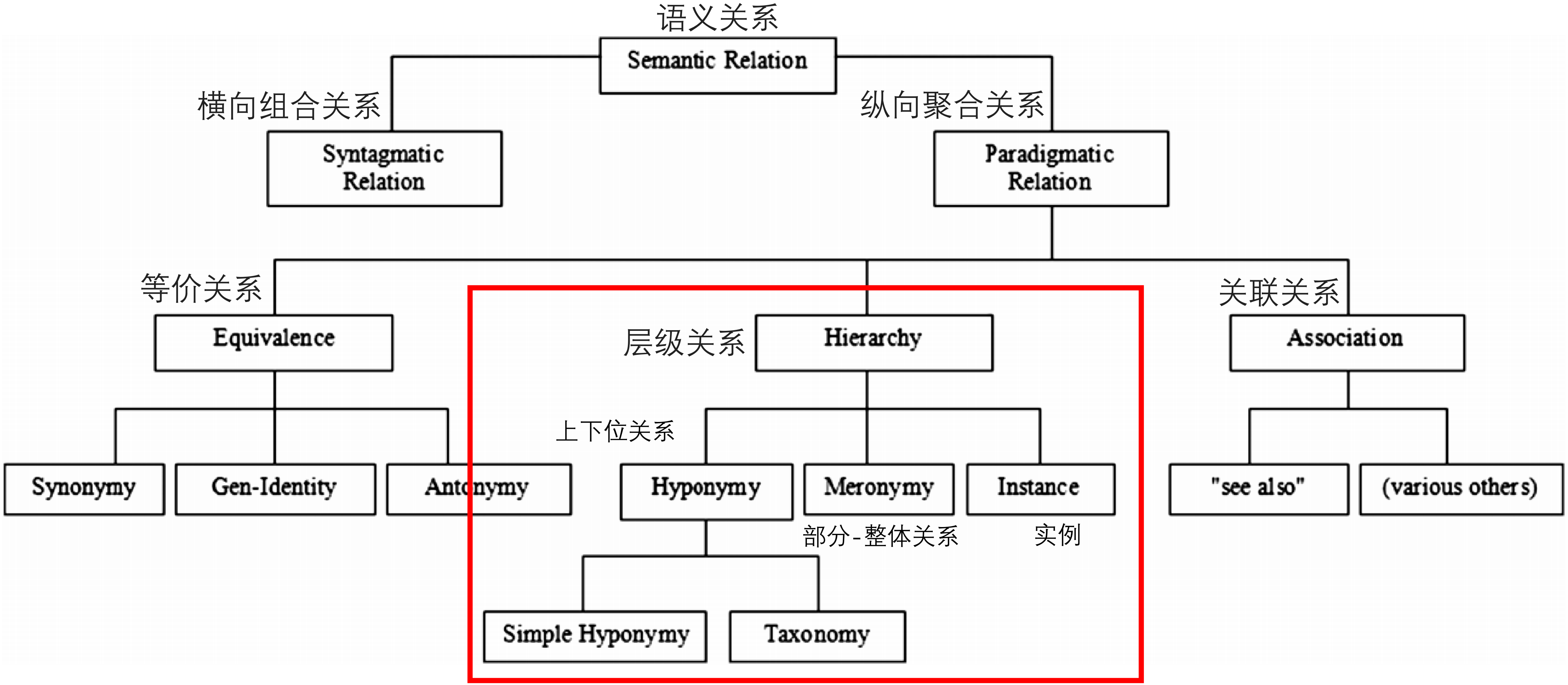


Figure 1. Ontology Learning Layer Cake

图片改编自 from the Figure 1 in Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini: 'Ontology Learning from Text: An Overview', 2003

情报学中语义关系的种类

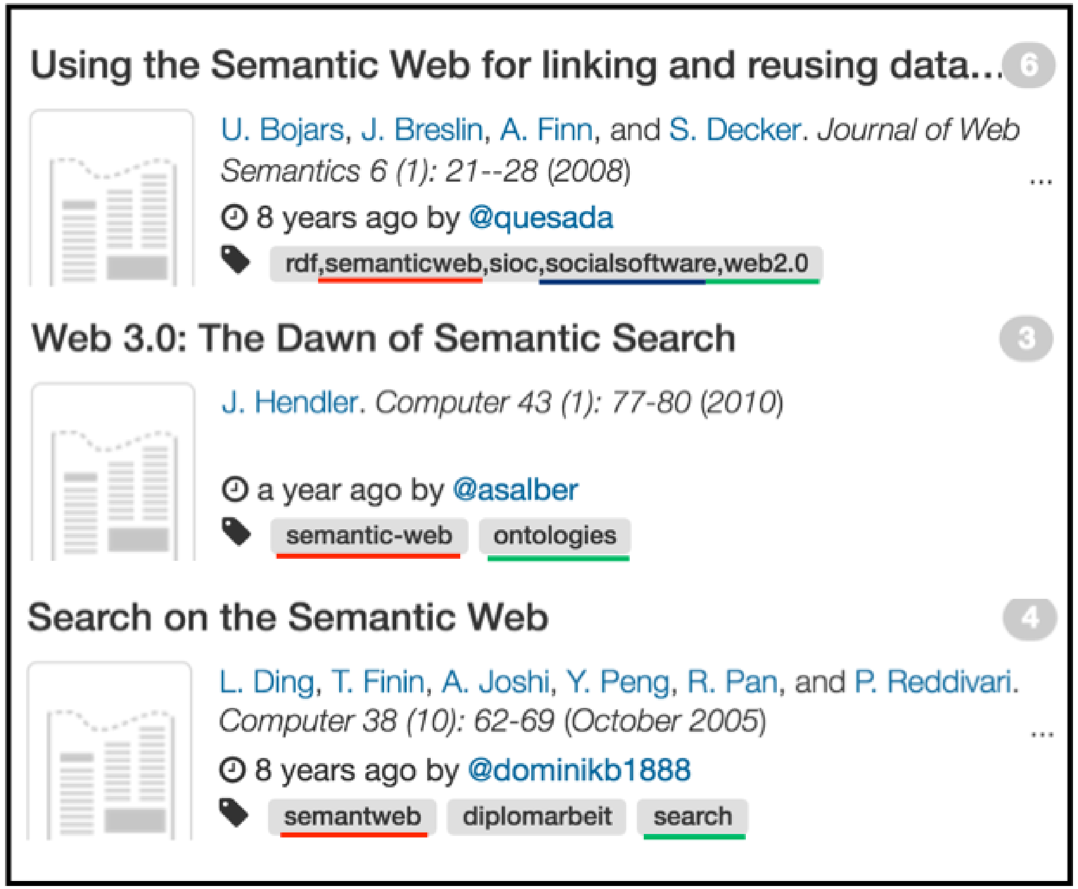


图片改编自: Stock, W. G. (2010). Concepts and semantic relations in information science. *Journal of the Association for Information Science and Technology*, *61*(10), 1951-1969.

概念抽取 Concept Extraction

- 词型: 通过词型来归一化
- 词义: 同义词的提取与合并; 多义词的词义消歧 (聚类)
- 外部资源: 匹配词到其他的词汇资源，比如维基百科

概念抽取: 词型归一化



Using the Semantic Web for linking and reusing data... 6

U. Bojars, J. Breslin, A. Finn, and S. Decker. *Journal of Web Semantics 6 (1): 21--28 (2008)* ...

8 years ago by @quesada

rdf,semanticweb,sioc,socialsoftware,web2.0

Web 3.0: The Dawn of Semantic Search 3

J. Hendler. *Computer 43 (1): 77-80 (2010)*

a year ago by @asalber

semantic-web ontologies

Search on the Semantic Web 4

L. Ding, T. Finin, A. Joshi, Y. Peng, R. Pan, and P. Reddivari. *Computer 38 (10): 62-69 (October 2005)* ...

8 years ago by @dominikb1888

semantweb diplomarbeit search

```
[en] Semantic_Web: semanticweb  semantic_web
SemanticWeb  Semantic_Web,_RDF_etc.  semantic+web
semanticWeb  #semanticweb  Semantic-Web  semantic_Web
semantic.web  semantic-web  semanticWeb,  semantic\\_web,
{SemanticWeb}  semanticweb,  semantic\\_web  Semantic_web
Semanticweb  semantweb  web:semanticweb  semantiweb
semanticwe  rdf,semanticweb,sioc,socialsoftware,web2.0
semantic_web,  {SemanticWeb  sematnic+web
[en] Social_Software: socialsoftware  social_software
SocialSoftware  social.software ...
```

```
[en] web2.0: Education,Web2.0,Pharmacy  Web2.0  "Web2.0"
web2.0  WEB2.0  web2.0, ...
[en] ontologies: ontologies  Ontologie  Ontologies
ontologie  ontologies,  Ontologies, ...
[en] search: SEARCH  1,search radar;search  Searching
sequences,search  processing;search  searching
searches,Library search ...
```

Dong, H., Wang, W., & Coenen, F. (2017). Deriving Dynamic Knowledge from Academic Social Tagging Data: A Novel Research Direction. In iConference 2017 Proceedings (pp. 661-666). https://doi.org/10.9776/17313

词表示: 用向量的方式表示标签

- 词-词向量，向量的维度是词汇数量
- **词-资源向量**，向量的维度是资源数量
- 词-用户向量，向量的维度是用户数量

	R1	R2	R3
news	1	0	0
Web2.0	1	1	1
knowledge	0	0	1

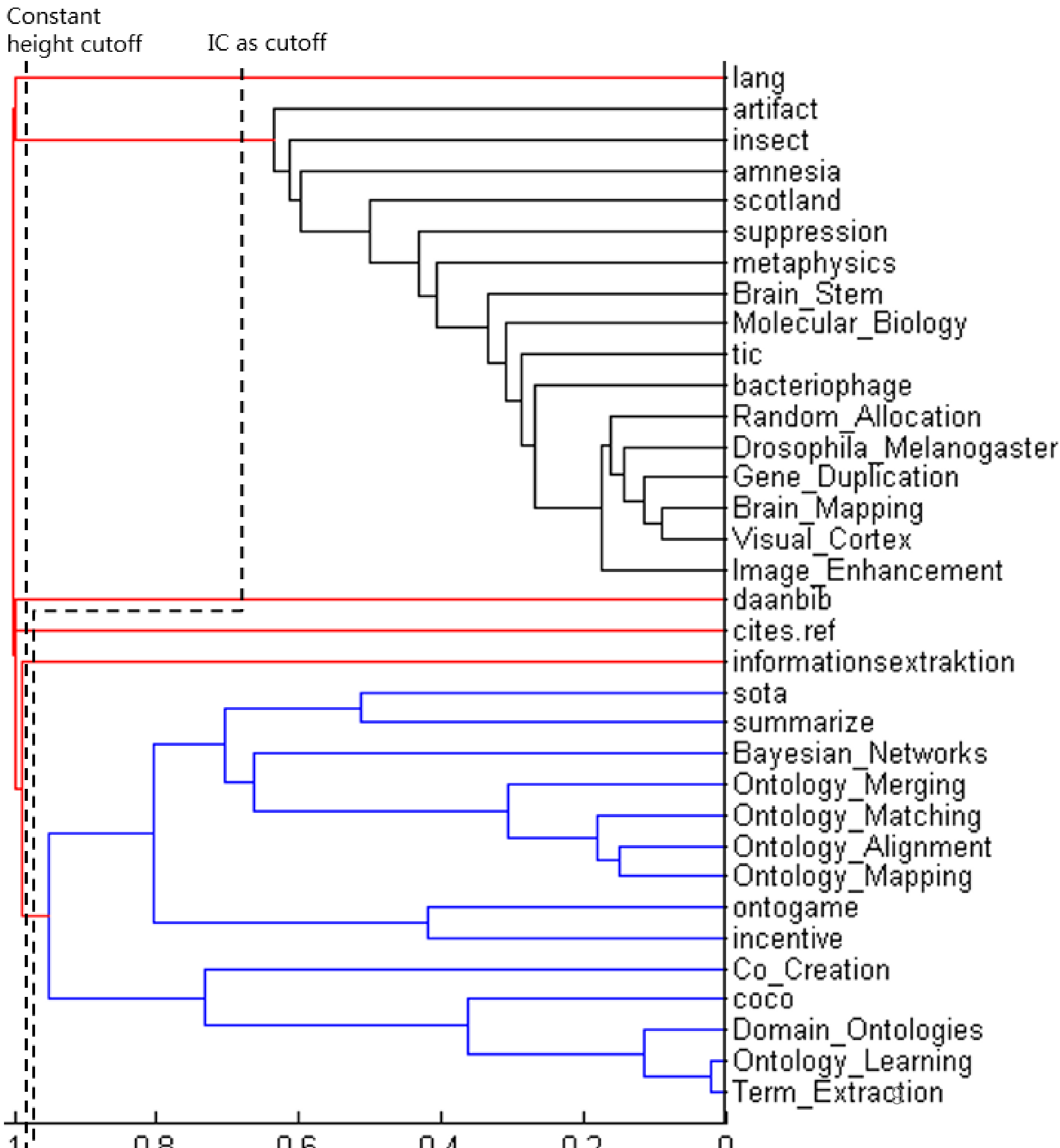
- 潜在语义表示 LSI (Latent Semantic Indexing)，设定向量维度
- 主题向量: LDA (Latent Dirichlet Allocation) Topic vector，设定向量维度
- 词嵌入: word2vec ， 设定向量维度，需要大量语料

概念抽取: 词聚类

将词表示成资源的向量，并进行降维

采用余弦距离计算相似度

使用分层聚类算法 (Chapter 8.3; Tan, Steinbach, & Kumar, 2006)



概念抽取：语义匹配

- 将标签匹配到现有的外部词表中

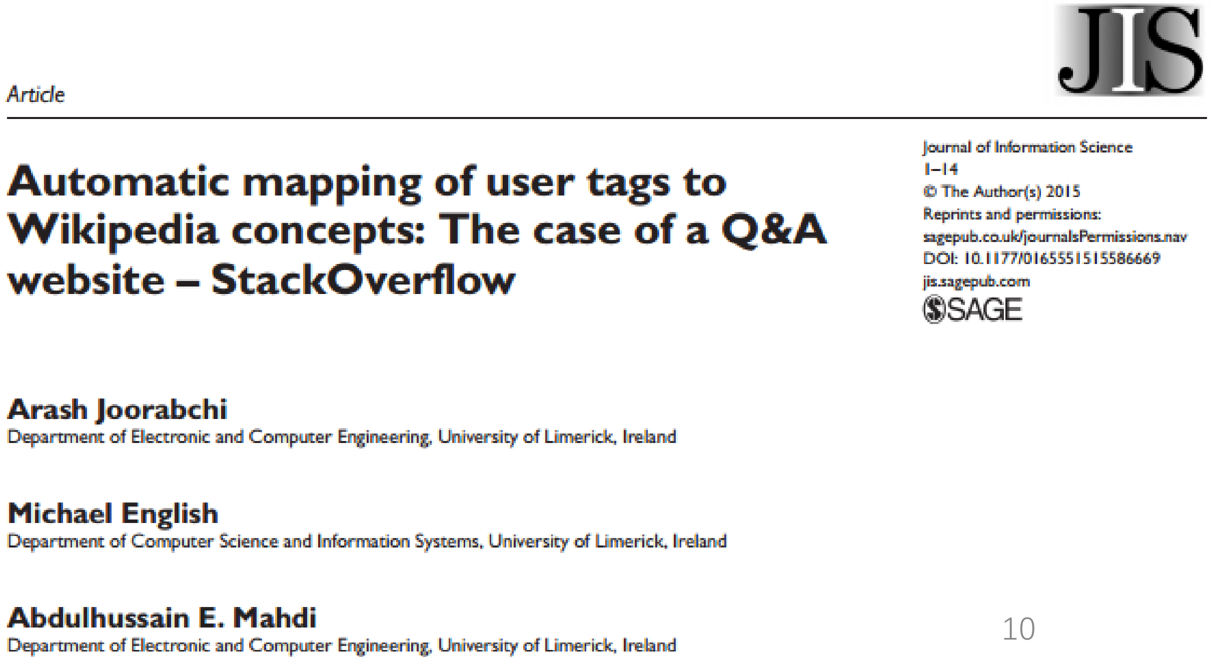
- 匹配到WordNet: 仅49%的标签可从语义上匹配到WordNet中 (Andrew, Pane & Zaihrayeu, 2011)

- 匹配到Wikipedia (Joorabchi, English, Mahdi, 2015)

- 匹配到以Dbpedia为主的

Linked Open Data Cloud

(García-Silva et al., 2015)



JIS

Article

Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow

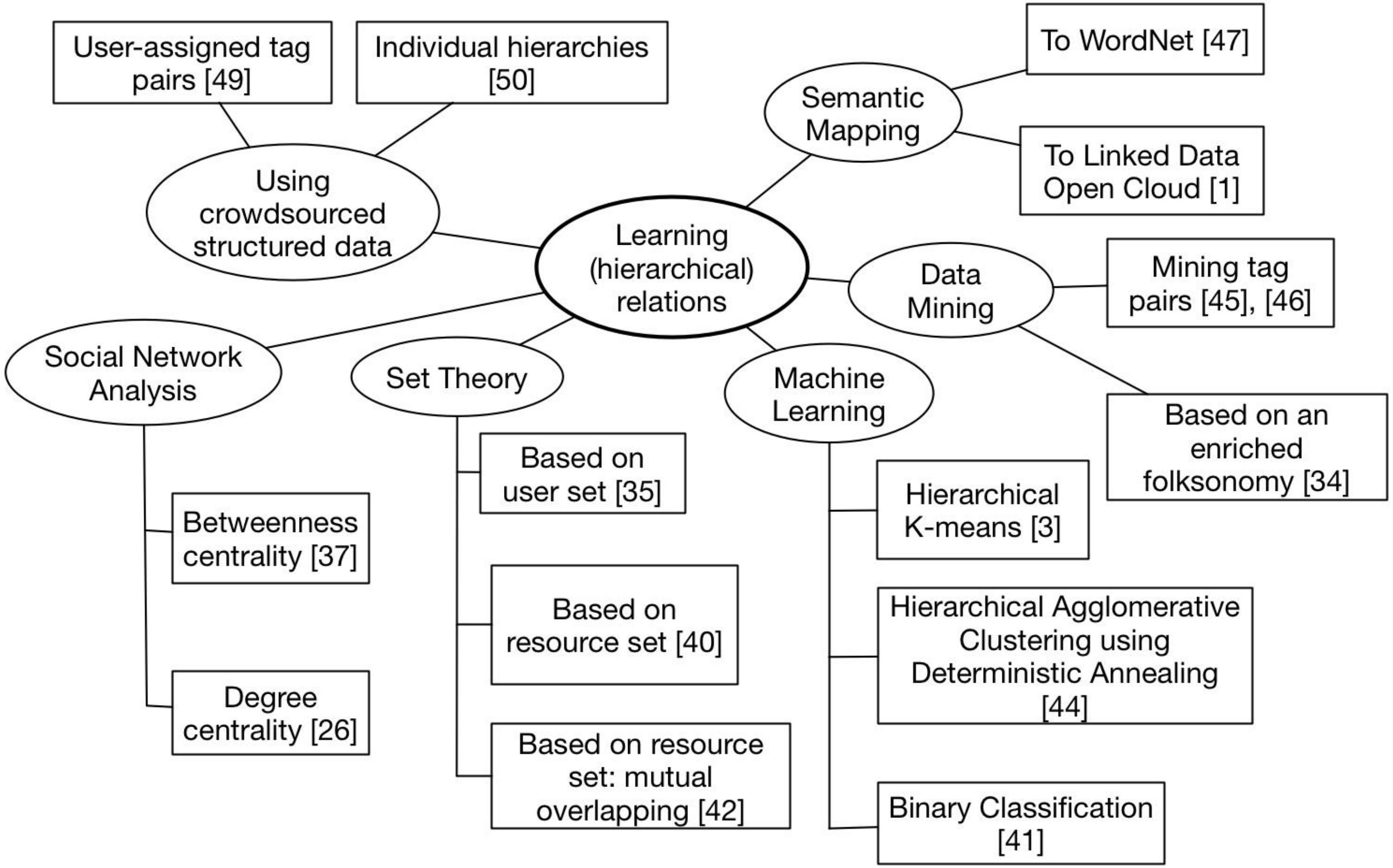
Journal of Information Science
1–14
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551515586669
jis.sagepub.com
SAGE

Arash Joorabchi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

Michael English
Department of Computer Science and Information Systems, University of Limerick, Ireland

Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

关系的形成 Relation Learning


User-assigned tag pairs [49]
Individual hierarchies [50]
Semantic Mapping
To WordNet [47]
To Linked Data Open Cloud [1]
Using crowdsourced structured data
Learning (hierarchical) relations
Data Mining
Mining tag pairs [45], [46]
Social Network Analysis
Set Theory
Machine Learning
Based on an enriched folksonomy [34]
Based on user set [35]
Betweenness centrality [37]
Hierarchical K-means [3]
Based on resource set [40]
Hierarchical Agglomerative Clustering using Deterministic Annealing [44]
Degree centrality [26]
Based on resource set: mutual overlapping [42]
Binary Classification [41]

H. Dong, W. Wang and H. N. Liang, "Learning Structured Knowledge from Social Tagging Data: A Critical Review of Methods and Techniques," *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Chengdu, 2015, pp. 307-314.

从标签中自动建立层级关系的主要方法

- 基于一定规则的方法
  - **社会网络分析**图中心性的方法 (Heymann, 2006)
  - 利用标签对应资源或用户的集合的包含度的方法 (Mika, 2005)
- 基于语义匹配的方法
  - 匹配到Dbpedia, WordNet, ConceptNet, Yago, ACM category, MeSH… (Strohmaier et al., 2012; García-Silva et al., 2015)
- 机器学习方法
  - 无监督方法: 分层聚类 (Strohmaier et al., 2012; Zhou et al., 2007)
  - 有监督方法: 提取特征进行二元分类 (Rêgo et al., 2015)

方法1: 基于社会网络分析的方法 (Heymann, 2006)

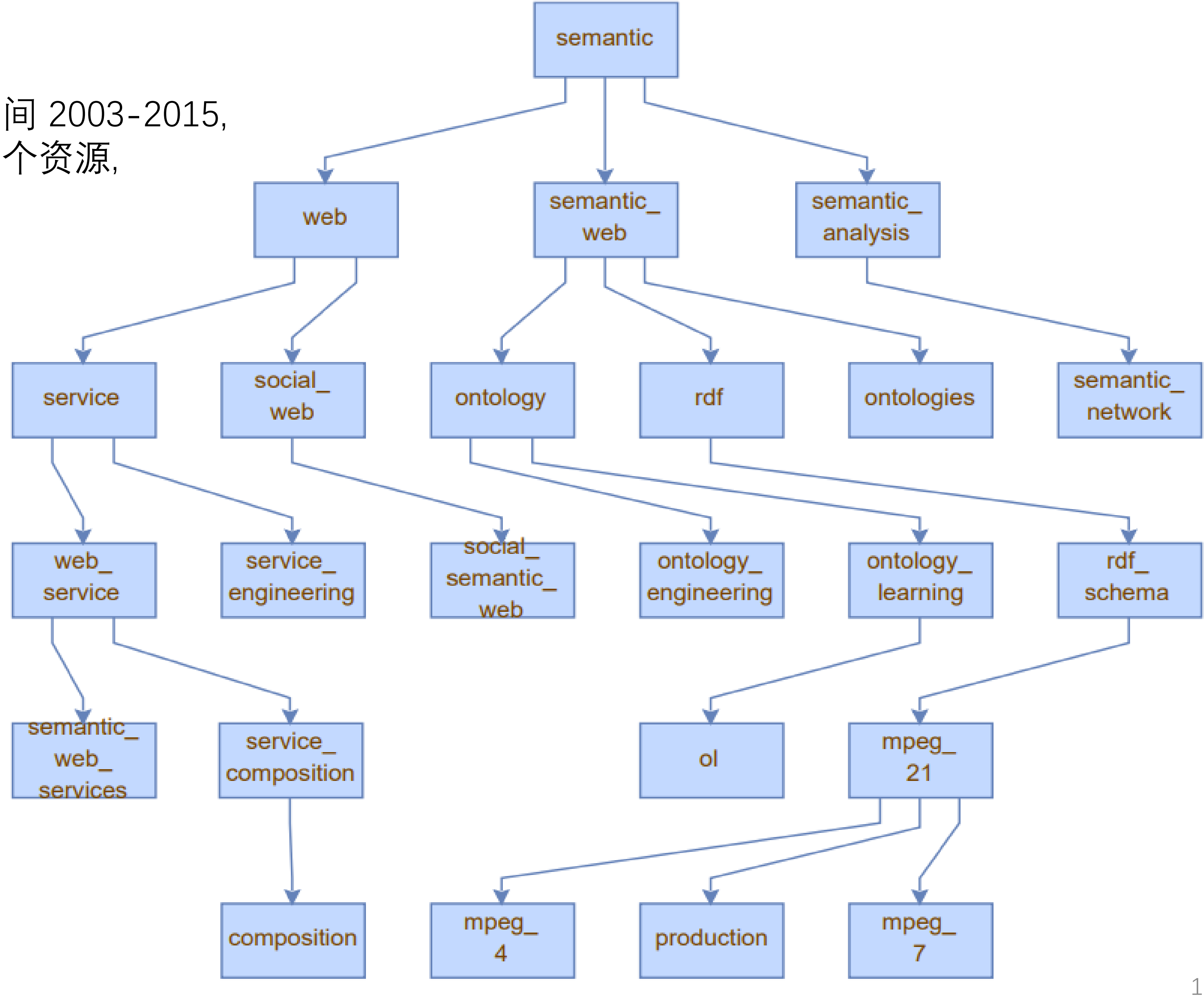
- 设想: 在标签相似度图中，有一个潜在的分类体系;
  中心性更高的标签，与其它标签连接更紧密的标签，含义更为宽泛

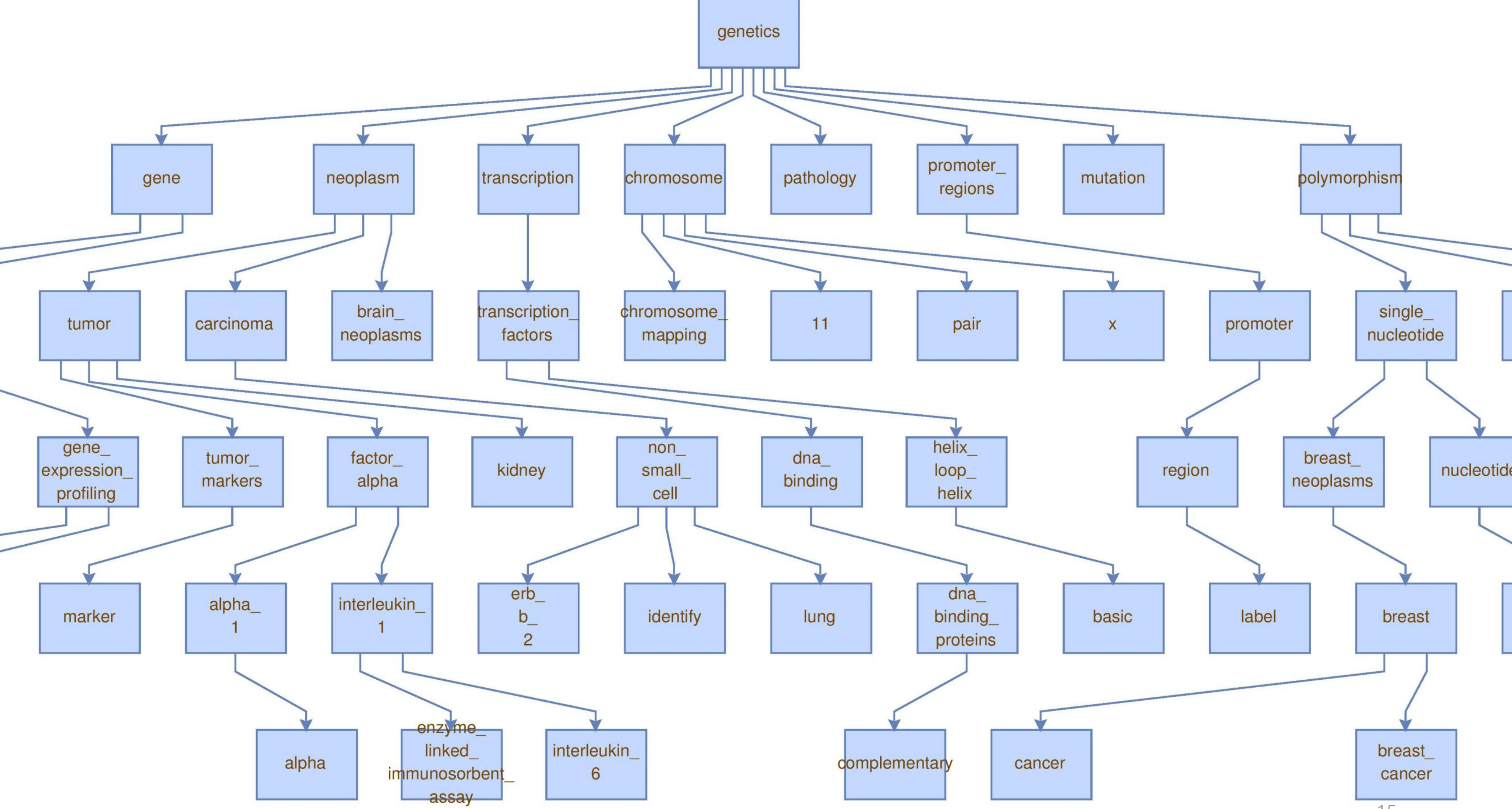
- 建立标签相似度无向图，将标签按照度中心性降序排列

- 从中心性最高的标签开始，依次添加到新的有向图中，将标签与图中的节点依次比较，若相似度大于某阈值，则列为该节点的下位类。

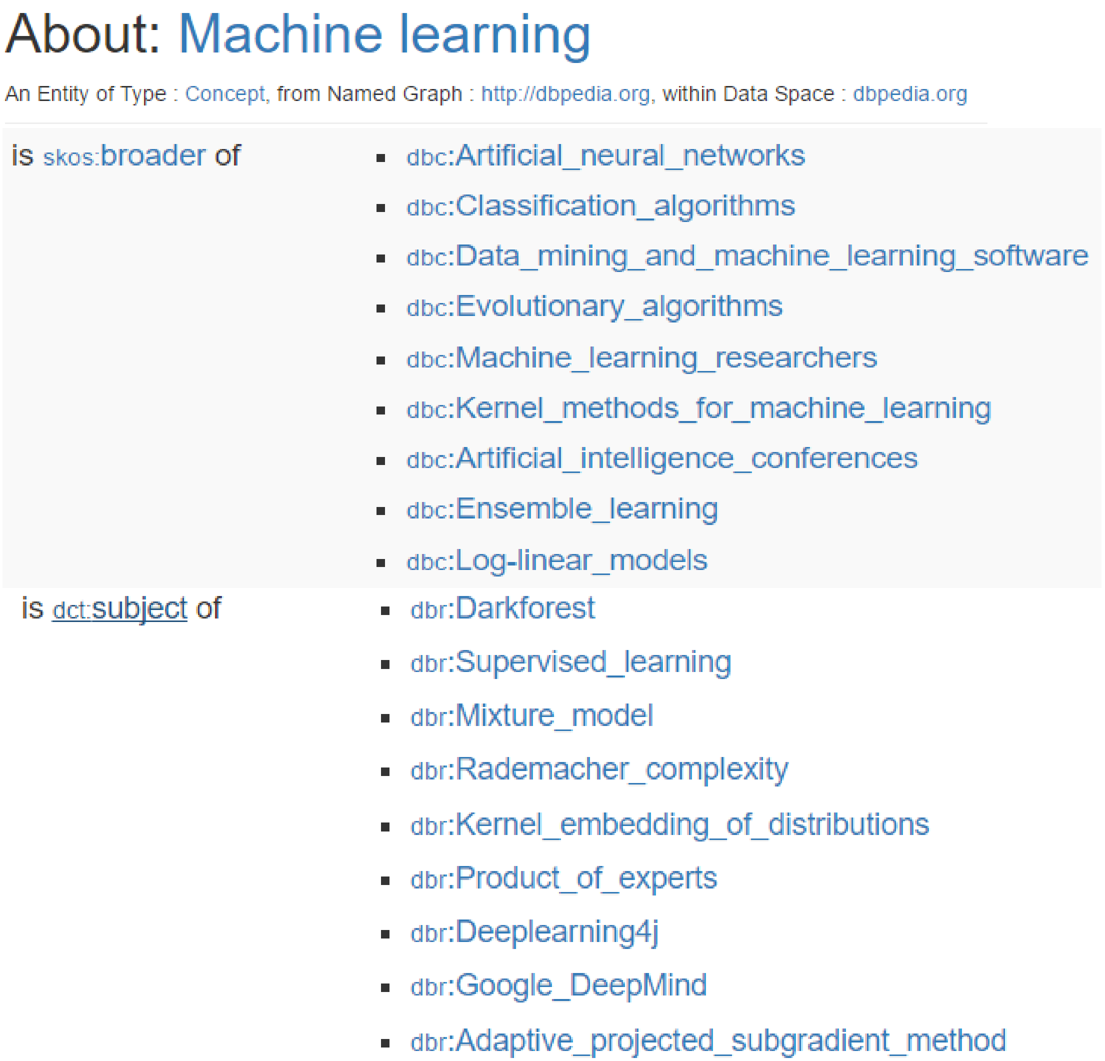
- 优点: 方法容易实现，不依赖外部资源
- 缺点: 建立的联系不完全正确，语义关系不明确

数据集: Bibsonomy dataset, 时间 2003-2015, 包括 3794882 个标注, 868015 个资源, 283858 个标签, 11103 个用户.




genetics
gene
neoplasm
transcription
chromosome
pathology
promoter_ regions
mutation
polymorphism
tumor
carcinoma
brain_ neoplasms
transcription_ factors
chromosome_ mapping
11
pair
x
promoter
single_ nucleotide
gene_ expression_ profiling
tumor_ markers
factor_ alpha
kidney
non_ small_ cell
dna_ binding
helix_ loop_ helix
region
breast_ neoplasms
nucleotide
marker
alpha_ 1
interleukin_ 1
erb_ b_ 2
identify
lung
dna_ binding_ proteins
basic
label
breast
alpha
enzyme_ linked_ immunosorbent_ assay
interleukin_ 6
complementary
cancer
breast_ cancer

方法2: 基于语义匹配的方法



About: Machine learning

An Entity of Type : Concept, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

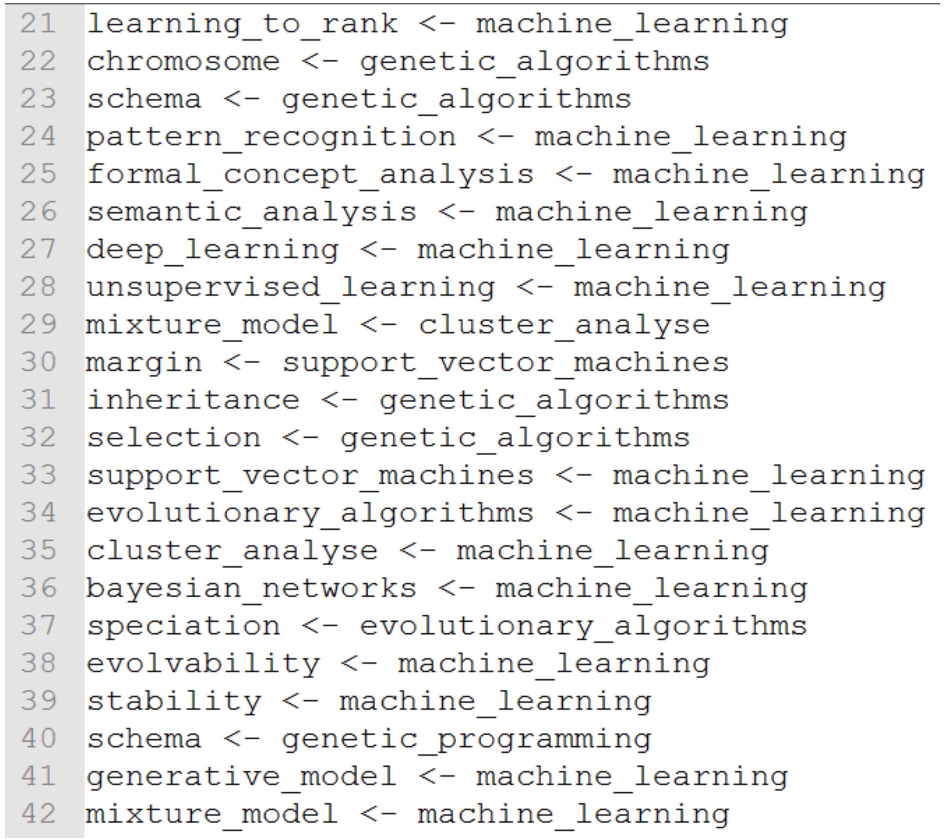
is skos:broader of

- dbc:Artificial_neural_networks
- dbc:Classification_algorithms
- dbc:Data_mining_and_machine_learning_software
- dbc:Evolutionary_algorithms
- dbc:Machine_learning_researchers
- dbc:Kernel_methods_for_machine_learning
- dbc:Artificial_intelligence_conferences
- dbc:Ensemble_learning
- dbc:Log-linear_models

is dct:subject of

- dbr:Darkforest
- dbr:Supervised_learning
- dbr:Mixture_model
- dbr:Rademacher_complexity
- dbr:Kernel_embedding_of_distributions
- dbr:Product_of_experts
- dbr:Deeplearning4j
- dbr:Google_DeepMind
- dbr:Adaptive_projected_subgradient_method

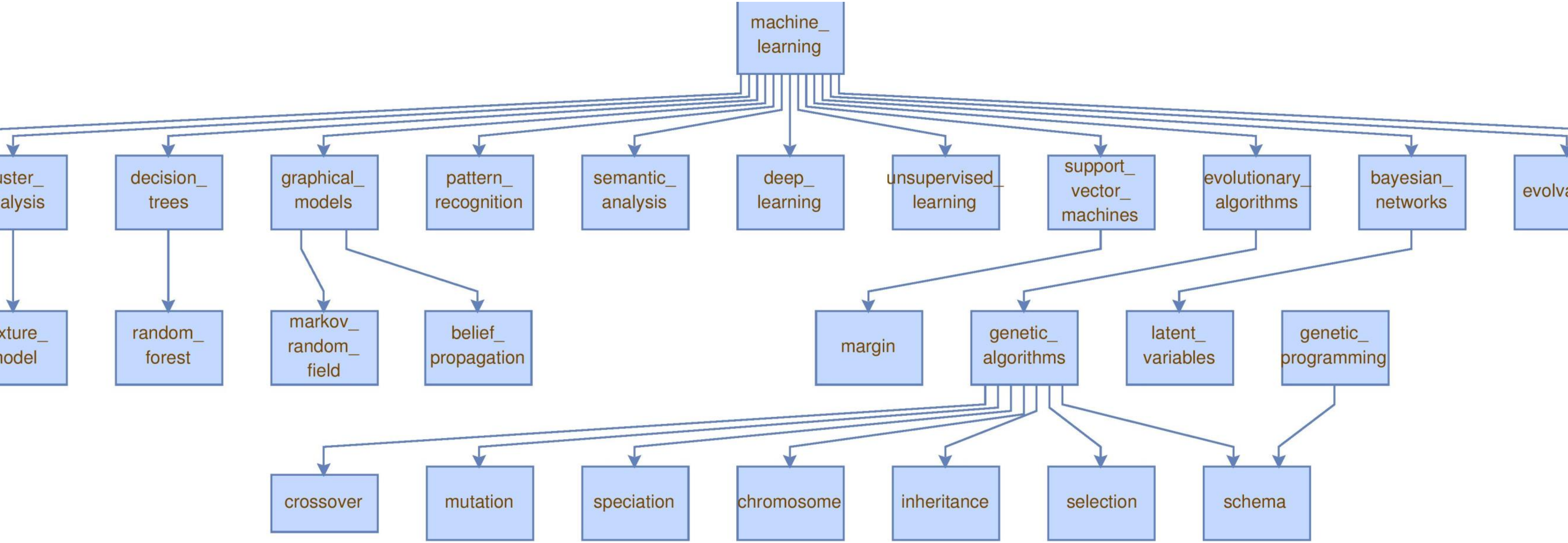
DBpedia concept pairs



```
learning_to_rank <- machine_learning
chromosome <- genetic_algorithms
schema <- genetic_algorithms
pattern_recognition <- machine_learning
formal_concept_analysis <- machine_learning
semantic_analysis <- machine_learning
deep_learning <- machine_learning
unsupervised_learning <- machine_learning
mixture_model <- cluster_analyse
margin <- support_vector_machines
inheritance <- genetic_algorithms
selection <- genetic_algorithms
support_vector_machines <- machine_learning
evolutionary_algorithms <- machine_learning
cluster_analyse <- machine_learning
bayesian_networks <- machine_learning
speciation <- evolutionary_algorithms
evolvability <- machine_learning
stability <- machine_learning
schema <- genetic_programming
generative_model <- machine_learning
mixture_model <- machine_learning
```

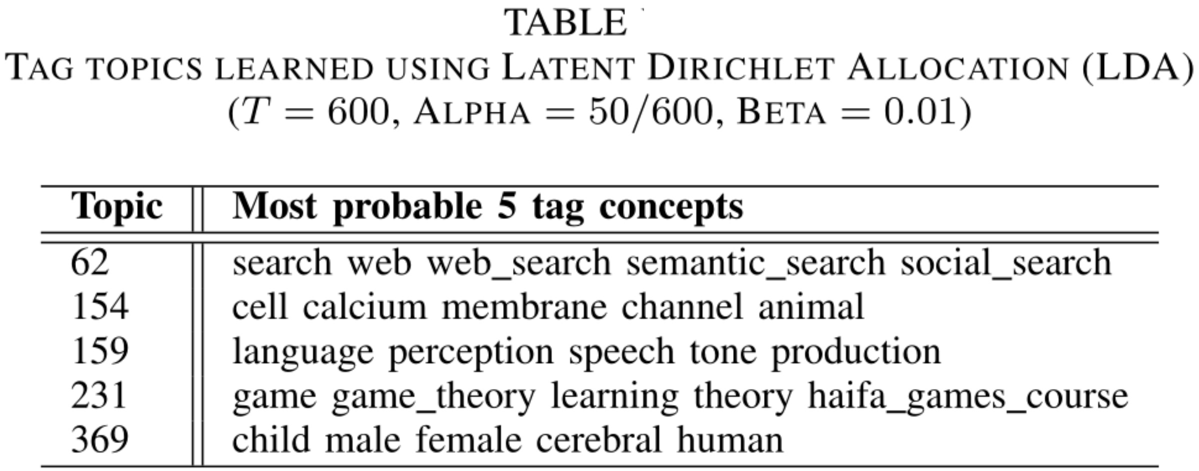
Matched tag concept pairs (positive data)

匹配机器学习下的类目



优点: 匹配到的关系有明确的语义, skos: broader, dct:subject
缺点: 依赖外部资源，受限于外部资源

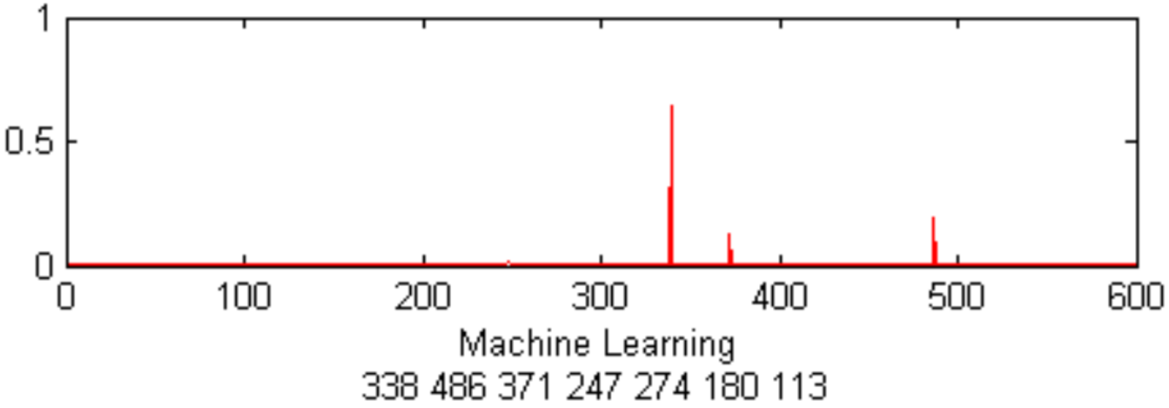
方法3: 基于主题模型的二元分类方法(实验中)

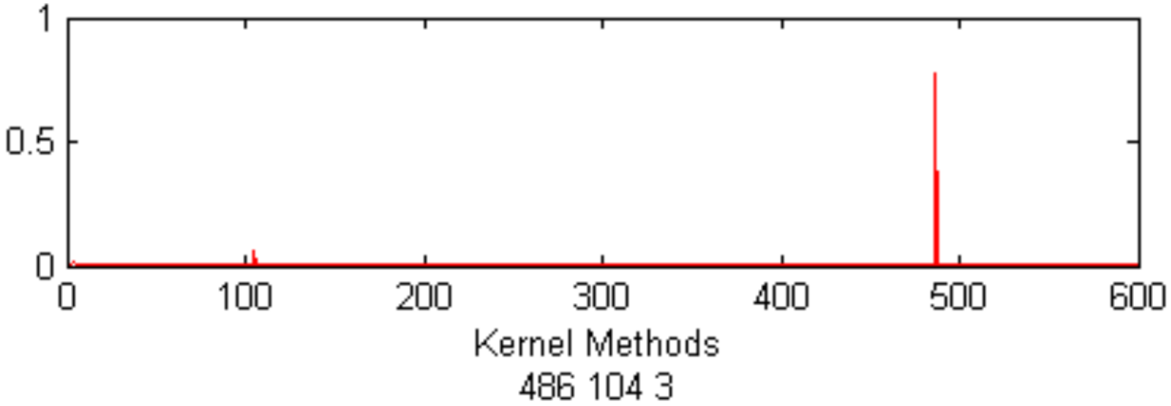


TABLE

TAG TOPICS LEARNED USING LATENT DIRICHLET ALLOCATION (LDA) ($T = 600$, ALPHA $= 50/600$, BETA $= 0.01$)

Topic	**Most probable 5 tag concepts**
62	search web web_search semantic_search social_search
154	cell calcium membrane channel animal
159	language perception speech tone production
231	game game_theory learning theory haifa_games_course
369	child male female cerebral human





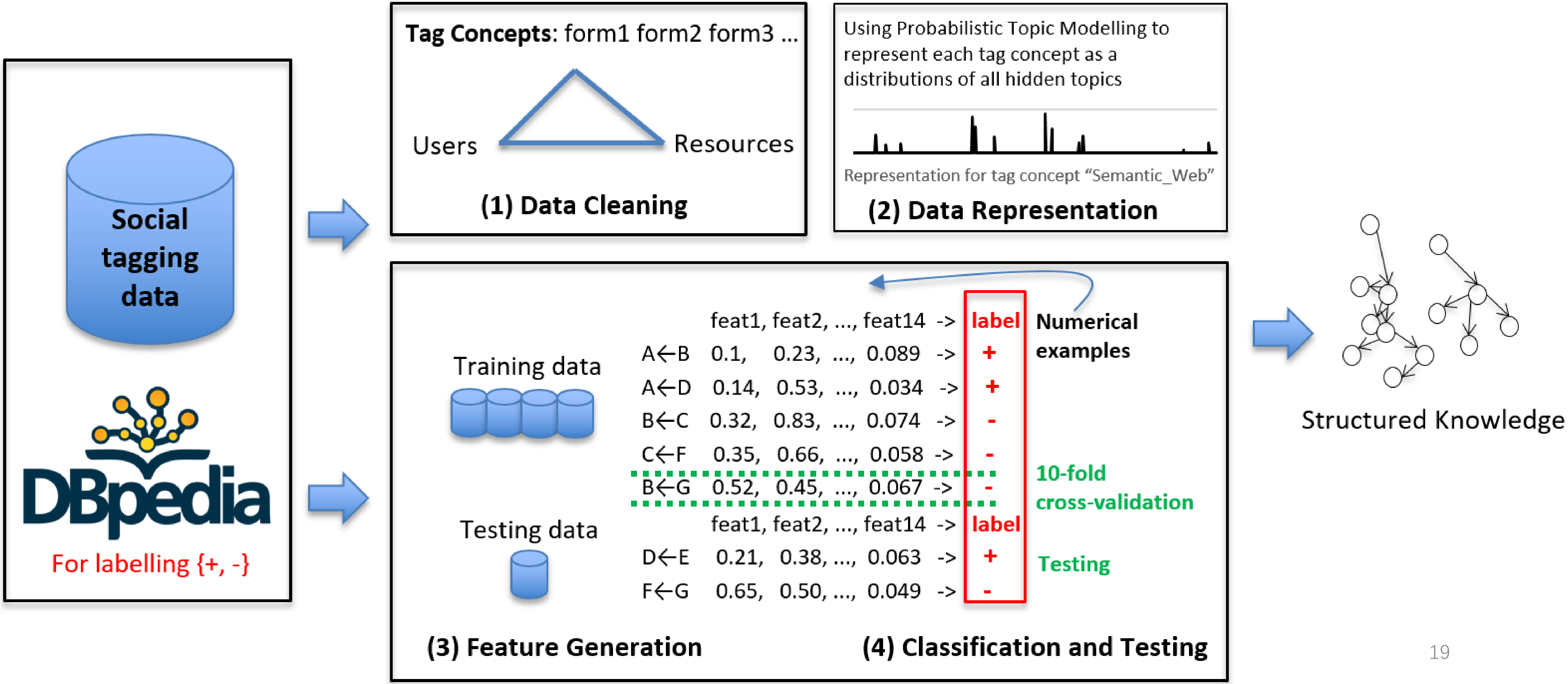
设想:

[1] 具有层次关系的标签必须有一定的相似度 (> p, p = 0.1)。

[2] 更显著地分布在多个主题的词汇，在含义上更为宽泛。

[3] 标签之间的层次关系与 边缘概率 p(A|B) 和 p(B|A) 相关。

基于主题模型的二元分类方法(实验中)



Social tagging data

DBpedia

For labelling {+, -}

Tag Concepts: form1 form2 form3 ...

Users

Resources

(1) Data Cleaning

Using Probabilistic Topic Modelling to represent each tag concept as a distributions of all hidden topics

Representation for tag concept “Semantic_Web”

(2) Data Representation

Training data

	feat1,	feat2,	...,	feat14	->	label
A←B	0.1,	0.23,	...,	0.089	->	+
A←D	0.14,	0.53,	...,	0.034	->	+
B←C	0.32,	0.83,	...,	0.074	->	-
C←F	0.35,	0.66,	...,	0.058	->	-
B←G	0.52,	0.45,	...,	0.067	->	-

Numerical examples

10-fold cross-validation

Testing data

	feat1,	feat2,	...,	feat14	->	label
D←E	0.21,	0.38,	...,	0.063	->	+
F←G	0.65,	0.50,	...,	0.049	->	-

Testing

(3) Feature Generation

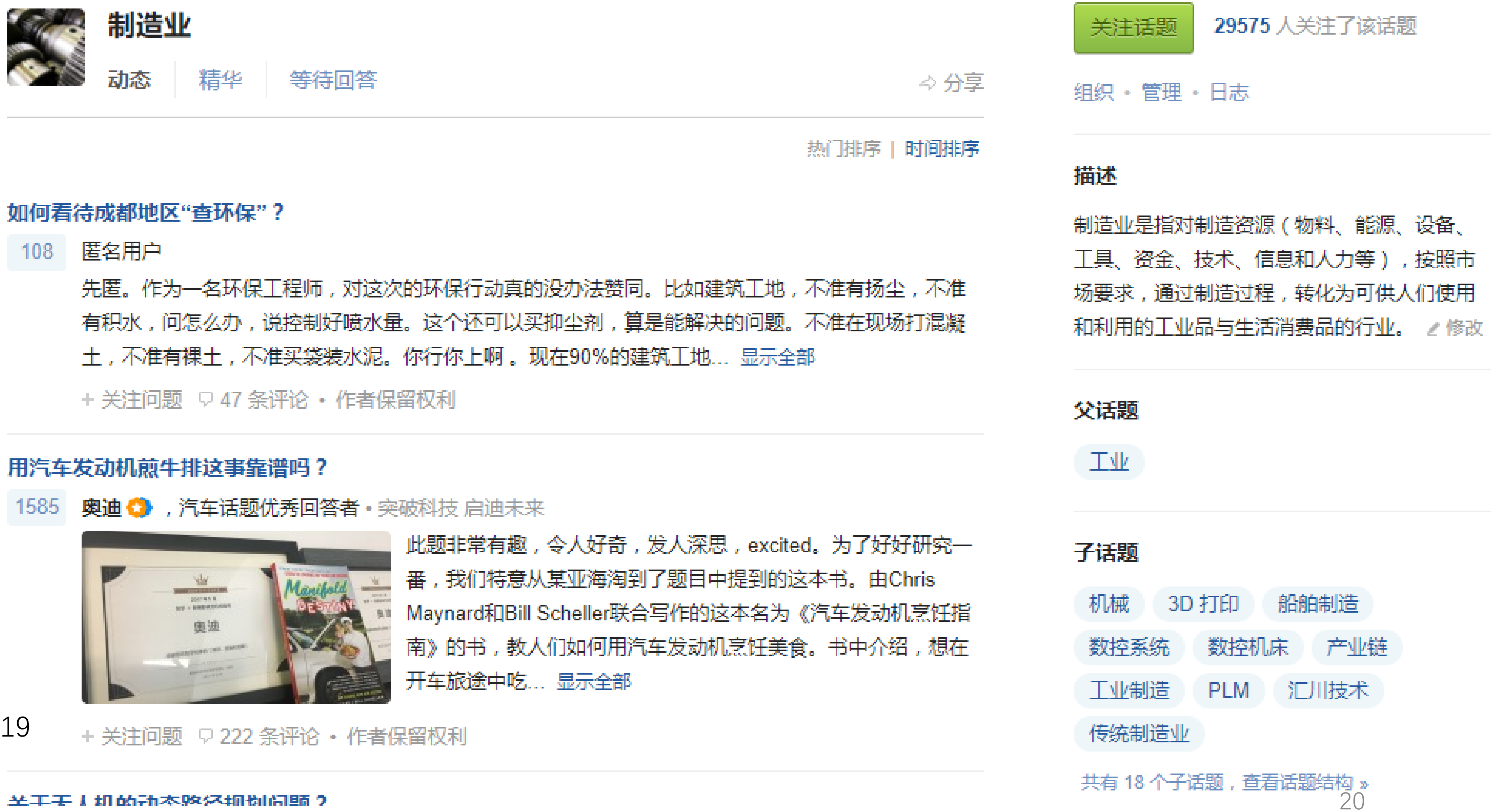
(4) Classification and Testing

Structured Knowledge

标签组织在系统中的运用

- 完善标签的导航，方便浏览资源

- 案例: 知乎、StackOverflow



制造业

动态 | 精华 | 等待回答 分享

热门排序 | 时间排序

如何看待成都地区“查环保”？

108 匿名用户

先匿。作为一名环保工程师，对这次的环保行动真的没办法赞同。比如建筑工地，不准有扬尘，不准有积水，问怎么办，说控制好喷水量。这个还可以买抑尘剂，算是能解决的问题。不准在现场打混凝土，不准有裸土，不准买袋装水泥。你行你上啊 。现在90%的建筑工地... 显示全部

+ 关注问题 47 条评论 • 作者保留权利

用汽车发动机煎牛排这事靠谱吗？

1585 奥迪 ，汽车话题优秀回答者 • 突破科技 启迪未来

此题非常有趣，令人好奇，发人深思，excited。为了好好研究一番，我们特意从某亚海淘到了题目中提到的这本书。由Chris Maynard和Bill Scheller联合写作的这本名为《汽车发动机烹饪指南》的书，教人们如何用汽车发动机烹饪美食。书中介绍，想在开车旅途中吃... 显示全部

+ 关注问题 222 条评论 • 作者保留权利

关于无人机的动态路径规划问题？

关注话题 29575 人关注了该话题

组织 • 管理 • 日志

描述

制造业是指对制造资源（物料、能源、设备、工具、资金、技术、信息和人力等），按照市场要求，通过制造过程，转化为可供人们使用和利用的工业品与生活消费品的行业。 修改

父话题

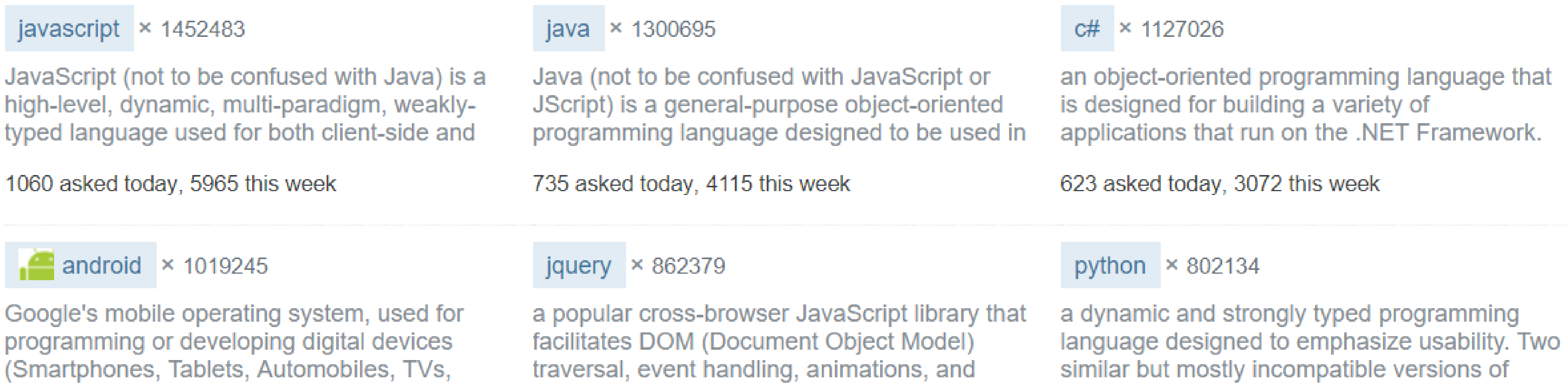
工业

子话题

机械 3D 打印 船舶制造 数控系统 数控机床 产业链 工业制造 PLM 汇川技术 传统制造业

共有 18 个子话题，查看话题结构 »

截图自: https://www.zhihu.com/topic/19551606/hot



javascript × 1452483

JavaScript (not to be confused with Java) is a high-level, dynamic, multi-paradigm, weakly-typed language used for both client-side and

1060 asked today, 5965 this week

java × 1300695

Java (not to be confused with JavaScript or JScript) is a general-purpose object-oriented programming language designed to be used in

735 asked today, 4115 this week

c# × 1127026

an object-oriented programming language that is designed for building a variety of applications that run on the .NET Framework.

623 asked today, 3072 this week

android × 1019245

Google's mobile operating system, used for programming or developing digital devices (Smartphones, Tablets, Automobiles, TVs,

620 asked today, 3714 this

jquery × 862379

a popular cross-browser JavaScript library that facilitates DOM (Document Object Model) traversal, event handling, animations, and

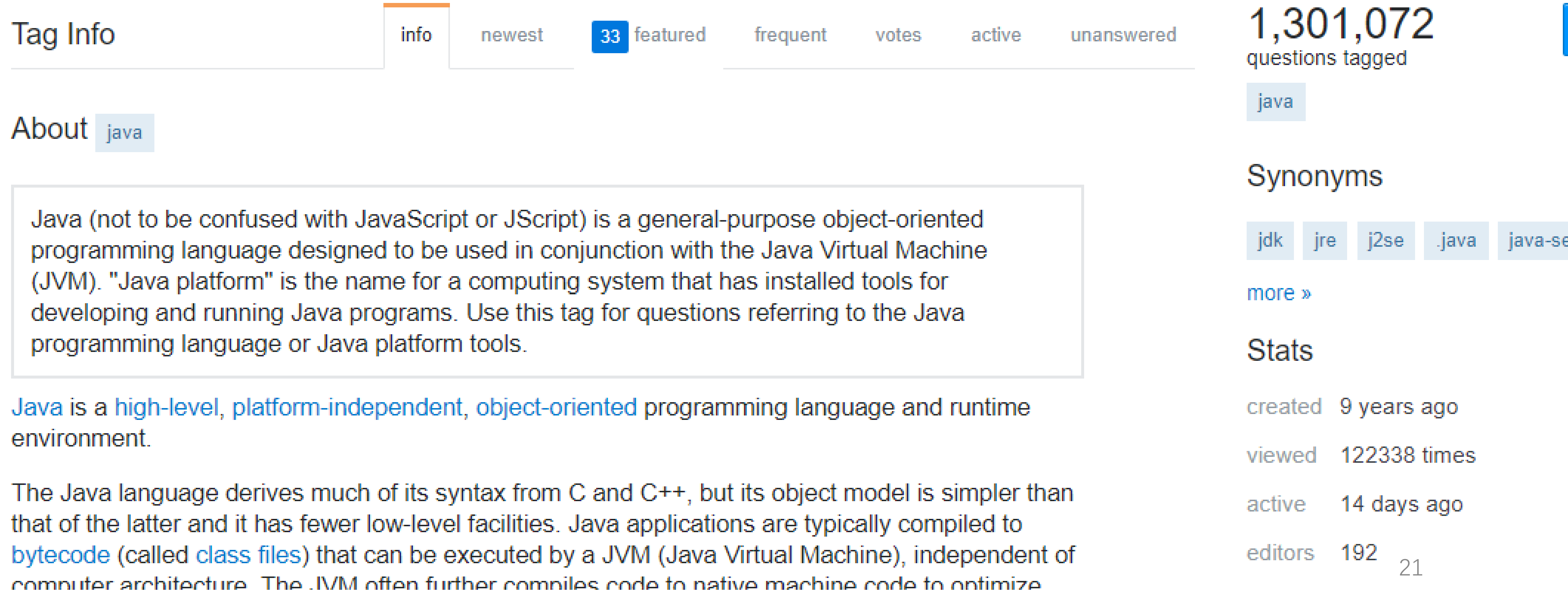
python × 802134

a dynamic and strongly typed programming language designed to emphasize usability. Two similar but mostly incompatible versions of

c++ × 528669

a general-purpose progra was originally designed a and keeps a similar synta

177 asked today, 1385 this



Tag Info

About java

Java (not to be confused with JavaScript or JScript) is a general-purpose object-oriented programming language designed to be used in conjunction with the Java Virtual Machine (JVM). "Java platform" is the name for a computing system that has installed tools for developing and running Java programs. Use this tag for questions referring to the Java programming language or Java platform tools.

Java is a high-level, platform-independent, object-oriented programming language and runtime environment.

The Java language derives much of its syntax from C and C++, but its object model is simpler than that of the latter and it has fewer low-level facilities. Java applications are typically compiled to bytecode (called class files) that can be executed by a JVM (Java Virtual Machine), independent of computer architecture. The JVM often further compiles code to native machine code to optimize

1,301,072 questions tagged java

Synonyms

jdk jre j2se .java java-se

more »

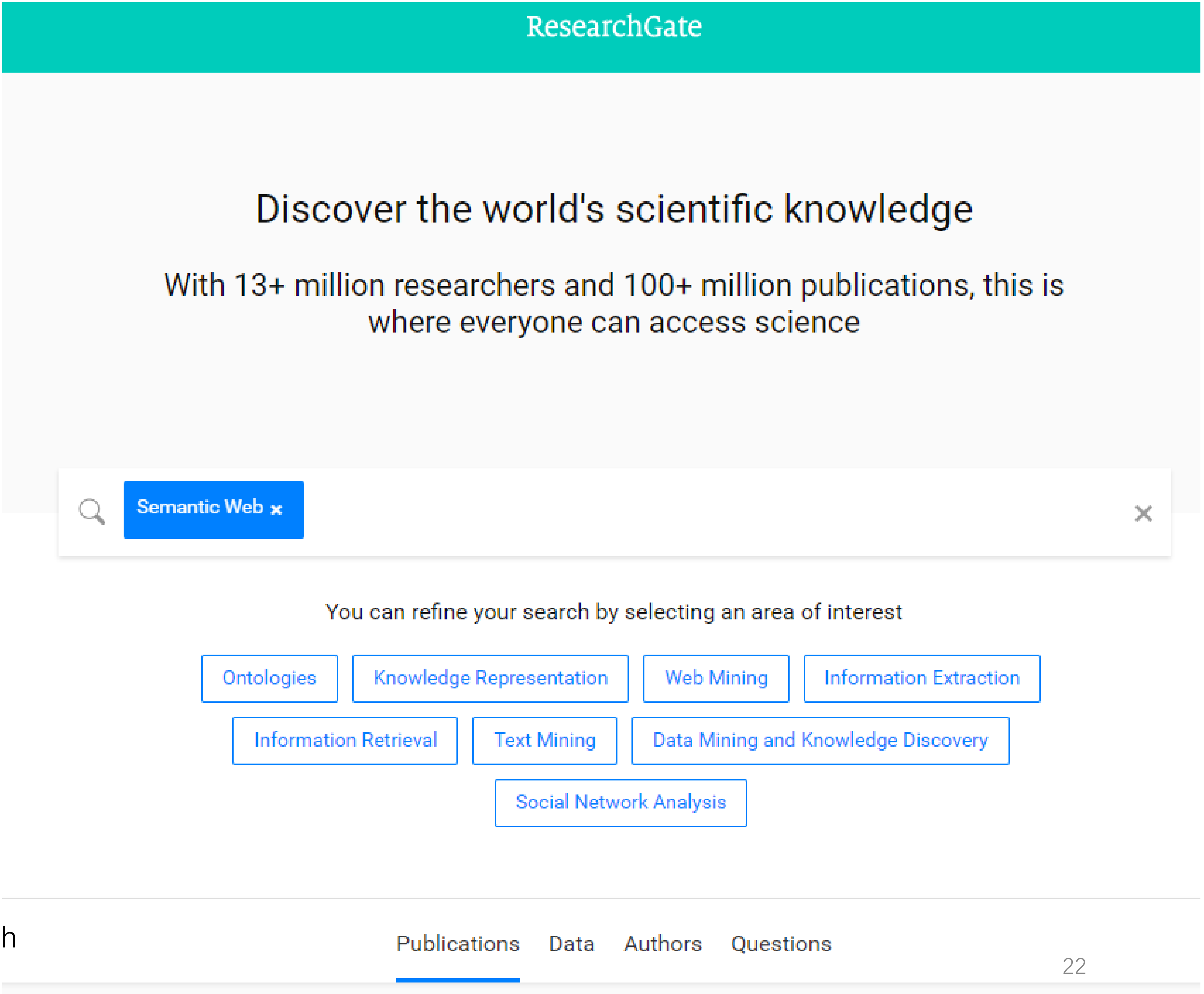
Stats

created	9 years ago
viewed	122338 times
active	14 days ago
editors	192

截图自: [1] https://stackoverflow.com/tags
[2] https://stackoverflow.com/tags/java/info

- 方便个性化的检索和推荐

(案例: ResearchGate 和豆瓣)



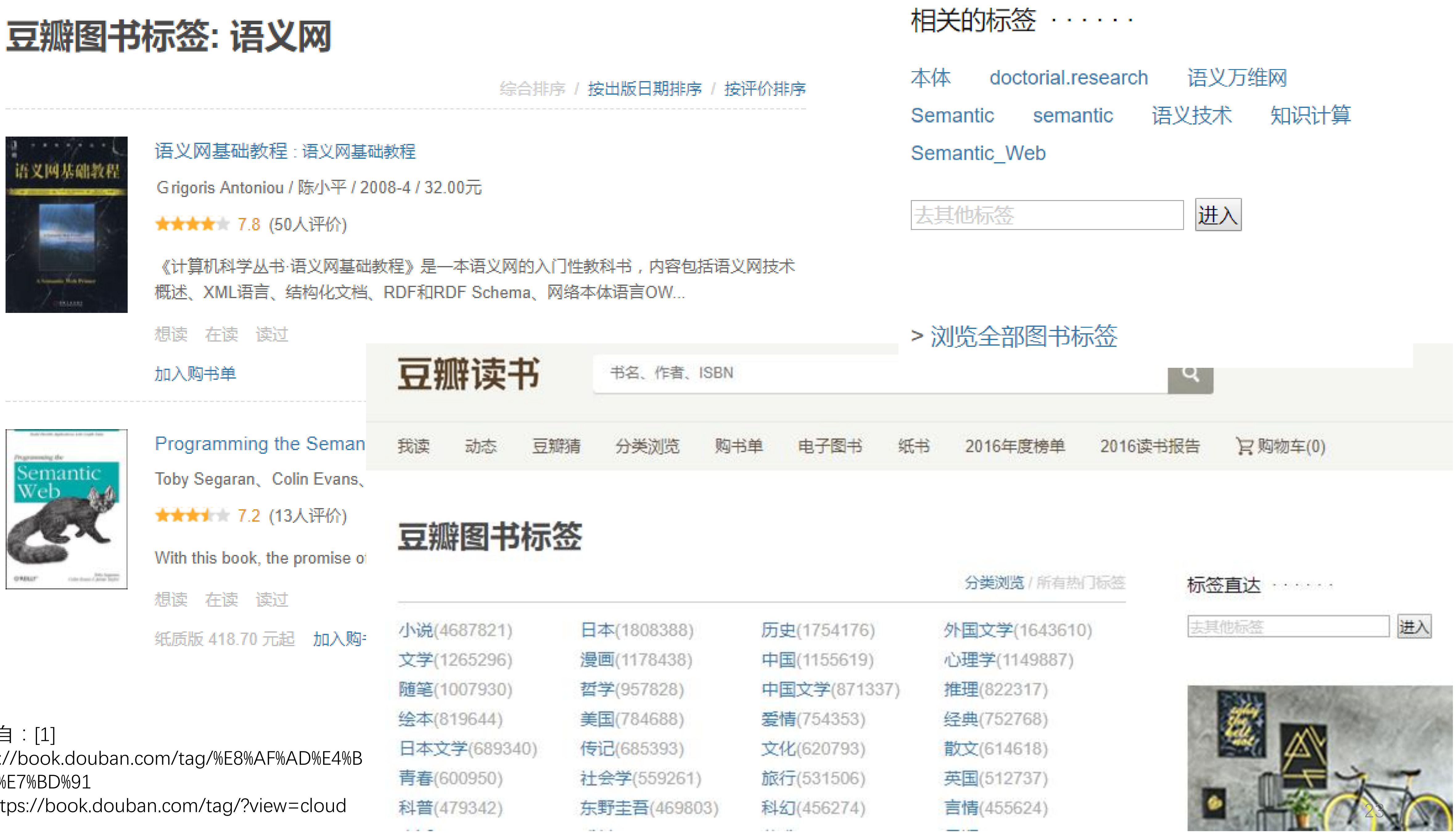
ResearchGate

Discover the world's scientific knowledge

With 13+ million researchers and 100+ million publications, this is where everyone can access science

You can refine your search by selecting an area of interest

截图自: https://www.researchgate.net/search



豆瓣图书标签: 语义网

综合排序 / 按出版日期排序 / 按评价排序

[语义网基础教程 : 语义网基础教程]

Grigoris Antoniou / 陈小平 / 2008-4 / 32.00元

★★★★☆ 7.8 (50人评价)

《计算机科学丛书·语义网基础教程》是一本语义网的入门性教科书，内容包括语义网技术概述、XML语言、结构化文档、RDF和RDF Schema、网络本体语言OW...

想读 在读 读过

加入购书单

Programming the Seman

Toby Segaran、Colin Evans、

★★★½☆ 7.2 (13人评价)

With this book, the promise o

想读 在读 读过

纸质版 418.70 元起 加入购

相关的标签 · · · · · ·

本体 doctorial.research 语义万维网
Semantic semantic 语义技术 知识计算
Semantic_Web

去其他标签 进入

> 浏览全部图书标签

豆瓣读书
书名、作者、ISBN
我读 动态 豆瓣猜 分类浏览 购书单 电子图书 纸书 2016年度榜单 2016读书报告 购物车(0)

豆瓣图书标签

分类浏览 / 所有热门标签

小说(4687821)	日本(1808388)	历史(1754176)	外国文学(1643610)
文学(1265296)	漫画(1178438)	中国(1155619)	心理学(1149887)
随笔(1007930)	哲学(957828)	中国文学(871337)	推理(822317)
绘本(819644)	美国(784688)	爱情(754353)	经典(752768)
日本文学(689340)	传记(685393)	文化(620793)	散文(614618)
青春(600950)	社会学(559261)	旅行(531506)	英国(512737)
科普(479342)	东野圭吾(469803)	科幻(456274)	言情(455624)

标签直达 · · · · · ·

去其他标签 进入

截图自：[1] https://book.douban.com/tag/%E8%AF%AD%E4%B9%89%E7%BD%91
[2] https://book.douban.com/tag/?view=cloud

总结

- Web 2.0时代的语义网需要对社会网络数据进行语义化的处理。
- 对巨量的社会标签进行有效组织依赖机器学习、自然语言处理、社会网络分析等方法。
- 从社会标签中抽取的概念和关系，可以用于完善系统的资源搜索、发现、推荐等功能。

参考文献

- Andrews, P., Pane, J., & Zaihrayeu, I. (2011). Semantic disambiguation in folksonomy: a case study. In *Advanced language technologies for digital libraries* (pp. 114-134). Springer, Berlin, Heidelberg.
- Dong, H., Wang, W., & Coenen, F. (2017). Deriving Dynamic Knowledge from Academic Social Tagging Data: A Novel Research Direction. In iConference 2017 Proceedings (pp. 661-666). https://doi.org/10.9776/17313
- Dong, H., Wang, W., & Liang, H. N. (2015, December). Learning Structured Knowledge from Social Tagging Data: A Critical Review of Methods and Techniques. In *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on* (pp. 307-314). IEEE.
- García-Silva, A., García-Castro, L. J., García, A., & Corcho, O. (2015). Building Domain Ontologies Out of Folksonomies and Linked Data. *International Journal on Artificial Intelligence Tools, 24(2).*
- Heymann, P., & Garcia-Molina, H. (2006). *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Retrieved from http://ilpubs.stanford.edu:8090/775/*
- Joorabchi, A., English, M., & Mahdi, A. E. (2015). Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow. *Journal of Information Science. doi:10.1177/0165551515586669*
- Rego, A. S. C, Marinho, L. B., & Pires, C. E. S. (2015). *A supervised learning approach to detect subsumption relations between tags in folksonomies. Paper presented at the Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain.*
- Souza, R. R., Tudhope, D., & Almeida, M. B. (2012). Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems. *Knowledge organization*, *39*(3), 179-192. Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini: 'Ontology Learning from Text: An Overview', 2003
- Stock, W. G. (2010). Concepts and semantic relations in information science. *Journal of the Association for Information Science and Technology*, *61*(10), 1951-1969.
- Strohmaier, M., Helic, D., Benz, D., K, C., #246, rner, & Kern, R. (2012). Evaluation of Folksonomy Induction Algorithms. *ACM Trans. Intell. Syst. Technol., 3(4), 1-22. doi:10.1145/2337542.2337559*
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining. Boston: Pearson Addison Wesley.*
- Zhou, M., Bao, S., Wu, X., & Yu, Y. (2007). *An unsupervised model for exploring hierarchical semantics from social annotations: Springer.*

谢谢聆听

董行 | hangdong@liverpool.ac.uk