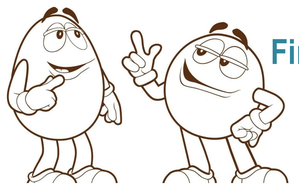
Using Full-text Academic Articles and Wikipedia to Find Alternative Free Bioinformatics Software

Shutian Ma; Chengzhi Zhang
Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094

What do we want to do?



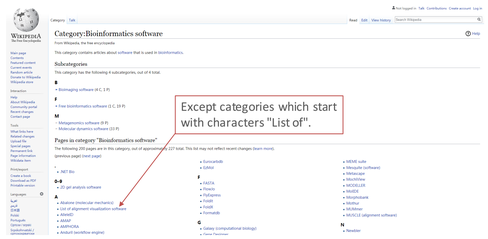
Find Free Software!

Abstract

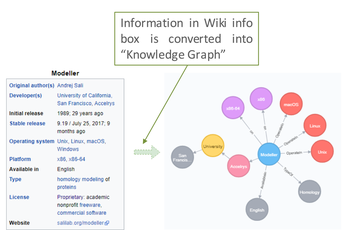
Taking bioinformatics software as a case study, this paper wants to find free software which are similar with commercial ones and have potential to be alternatives. Content and network information are applied for preference-oriented results, which encapsulates similarity in how people describe them in wiki and how people use them in research.

Method

Find Bioinformatics Software Information that Tells us Which is Free/Commercial



Represent Software Using Full-text and Wiki

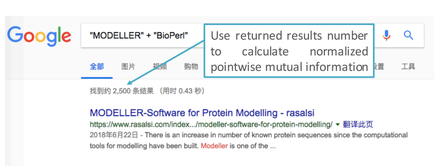


Compute Similarity between Free and Commercial Software

when doing linear combinations, all weights are set to be one.

Oriented Preference	Similarity Combination Description
Wiki-Content	Linear weighted sum of similarities in calculated based on vectors learned by LSI, LDA and Doc2Vec
Paper-Content	Linear weighted sum of similarities in calculated based on vectors learned by LSI, LDA and Doc2Vec
Wiki-Graph	Linear weighted sum of similarities calculated based on vectors learned by Node2Vec, Differ2Vec and struc2vec
Wiki-All	Linear weighted sum of Wiki-Content and Wiki-Graph similarities
PaperContent+WikiContent	Linear weighted sum of Paper-Content and Wiki Content similarities
PaperContent+WikiGraph	Linear weighted sum of Paper-Content and Wiki Graph similarities
Wiki+Paper	Linear weighted sum of Wiki-All and Paper-Content similarities

Construct Ground Truth and Evaluate Recommendation Result



❖ Software Representation Generation

LSI, LDA and Doc2Vec – Represent software via vectors based on content in 100 dimensions

Node2Vec, Differ2Vec and struc2vec – Represent software via vectors based on graph in 128 dimensions

❖ Software Graph Construction

Node Type	Value
developed by what kind of team	university, company and person
year of stable release	14 different years
written in what kind of programming language	17 languages
operation system	Linux, Unix, Windows and MacOS
applied platform	6 kinds
available language	English or cross-language
software type	44 types

Data and Tools

❖ Scientific literature

114,510 papers in XML format - PLOS ONE

11,013 articles containing bioinformatics software

❖ Software list

143 specific bioinformatics software - Wiki

20 commercial software according to their licenses

Only 97 software have info box information

❖ Tools

LSI and Doc2Vec - Genism

Python package of LDA

Node2Vec - OpenNE

Differ2Vec and struc2vec - Github

Experimental Result

Oriented Preference	nDCG@15	nDCG@20	nDCG@100	MAP@15	MAP@20	MAP@100	P@15	R@15
Wiki-Content	0.4312	0.4512	0.6997	0.1217	0.1624	0.8084	0.9933	0.1220
Paper-Content	0.3999	0.4237	0.6901	0.1187	0.1584	0.8036	0.9767	0.1198
Wiki-Graph	0.4112	0.4456	0.5258	0.1013	0.1369	0.4982	0.8933	0.1098
Wiki-All	0.4654	0.4915	0.7103	0.1228	0.1638	0.8099	1.0000	0.1228
PaperContent+WikiContent	0.4038	0.4219	0.6886	0.1197	0.1593	0.8048	0.9833	0.1207
PaperContent+WikiGraph	0.4313	0.4565	0.5293	0.1219	0.1622	0.5471	0.9933	0.122
Wiki+Paper	0.4268	0.447	0.7001	0.1223	0.1626	0.8094	0.9967	0.1224

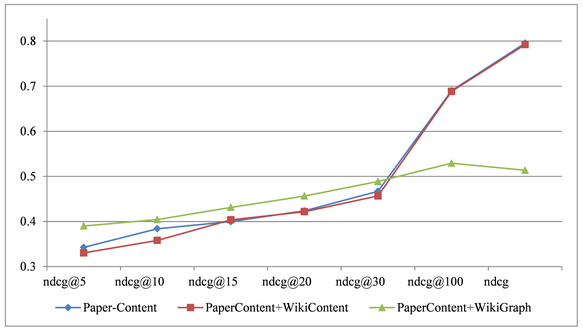
☑ Use Wiki content ☐ Use full-text paper content

☑ Wiki content ☐ Wiki graph

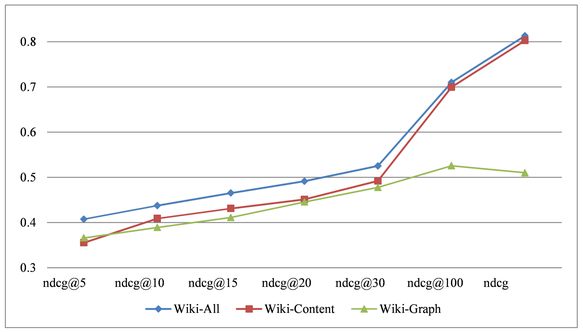
☑ Combine Wiki content & graph ☐ Wiki content ☐ Wiki graph

☑ Graph embedding can help to improve Paper-based recommendation.

☑ Combine all information, recommendation performance isn't getting much higher as expected.



Paper-based Software Recommendation Performance



Wiki-based Software Recommendation Performance

Conclusion

- ✓ Wikipedia content and info box can be balanced together for an efficient software recommendation technique.
- ✓ Graph-based information can help to rich semantic information.
- ✓ It's not suitable to use such kind of full-text publication data set to represent entities like software or some others in research

Contact Information

Shutian Ma: mashutian0608@hotmail.com
Chengzhi Zhang: zhangcz@njust.edu.cn