MONASH University

Tell Them Apart:

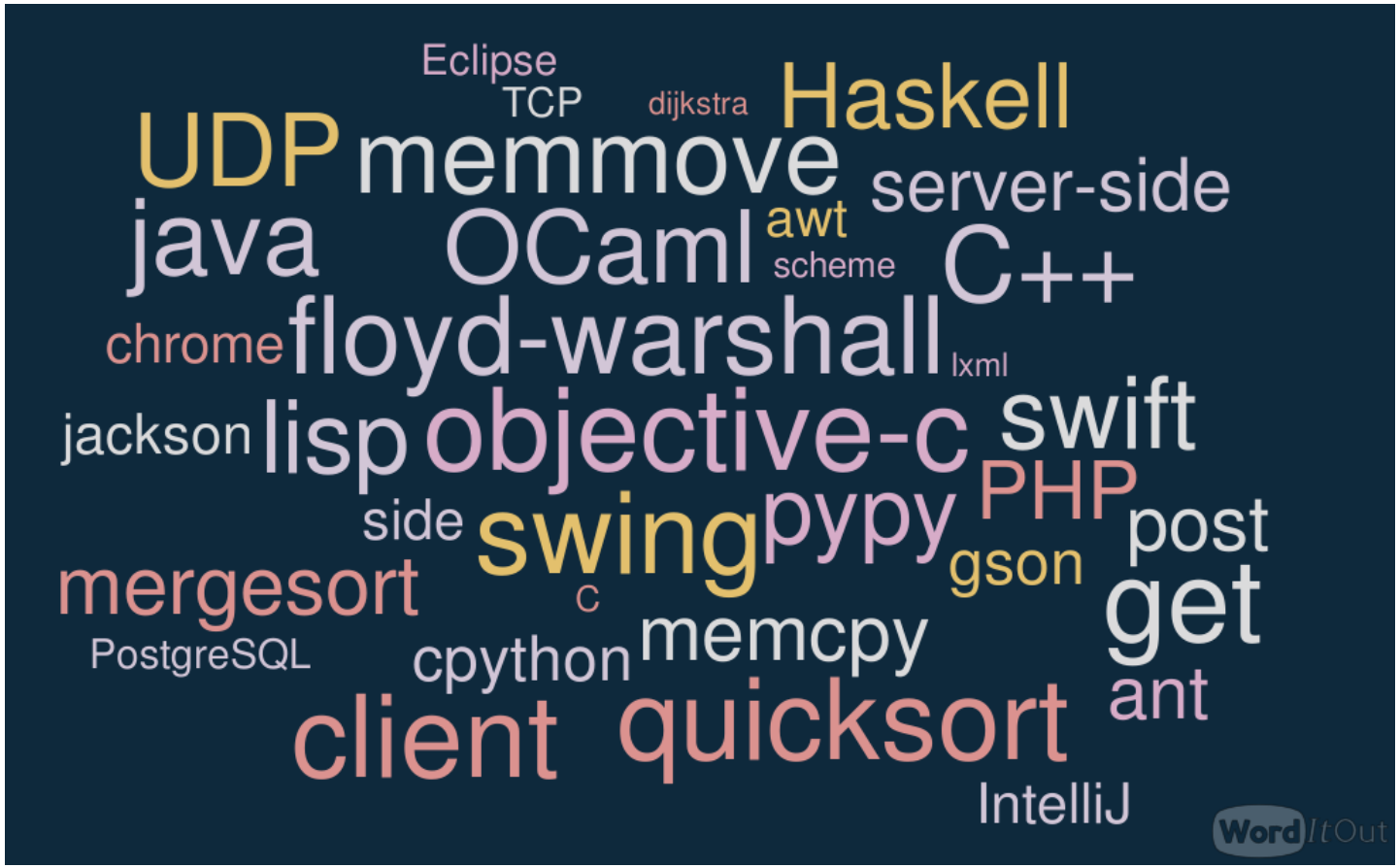
Distilling Technology Differences from Crow-Scale Comparison Discussions

Huang, Yi, Chunyang Chen, Zhenchang Xing, Tian Lin, and Yang Liu. "Tell them apart: distilling technology differences from crowd-scale comparison discussions." In *ASE*, pp. 214-224. 2018.

Tell Them Apart:

Distilling Technology Differences from Crow-Scale Comparison Discussions

How can we help developers make an informed choice when comparing alternative technologies?


Eclipse
TCP
dijkstra
Haskell
UDP
memmove
server-side
java
OCaml
awt
scheme
C++
chrome
floyd-warshall
lxml
jackson
lisp
objective-c
swift
side
swing
pypy
PHP
post
gson
mergesort
c
get
PostgreSQL
cpython
memcpy
ant
client
quicksort
IntelliJ
WordItOut

Java or Python?

Eclipse or Intellij?

POST or GET?

AWT or Swing?

MySQL or PostgreSQL?

Quicksort or Merge sort?

- Chen, Chunyang, Sa Gao, and Zhenchang Xing. "Mining analogical libraries in q&a discussions--incorporating relational and categorical knowledge into word embedding." In *2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER)*, vol. 1, pp. 338-348. IEEE, 2016.
- Chen, Chunyang, and Zhenchang Xing. "Similartech: automatically recommend analogical libraries across different programming languages." In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 834-839. IEEE, 2016.
- Chen, Chunyang, Zhenchang Xing, and Yang Liu. "What's spain's paris? mining analogical libraries from q&a discussions." *Empirical Software Engineering* 24, no. 3 (2019): 1155-1194.
- Chen, Chunyang, Zhenchang Xing, Yang Liu, and Kent Long Xiong Ong. "Mining likely analogical apis across third-party libraries via large-scale unsupervised api semantics embedding." *IEEE Transactions on Software Engineering* (2019).

Current Solutions

1. **Try them out**
    - Time-consuming
    - Labour expensive

Database
- MariaDB
- PostgreSQL
- SQL Server
- MySQL
- …

Library
- NLTK
- Stanford NLP
- OpenNLP
- SpaCy
- …

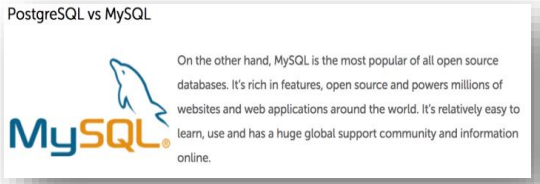
Sort Algorithms
- Bubble sort
- Selection sort
- Quicksort
- Merge sort
- …

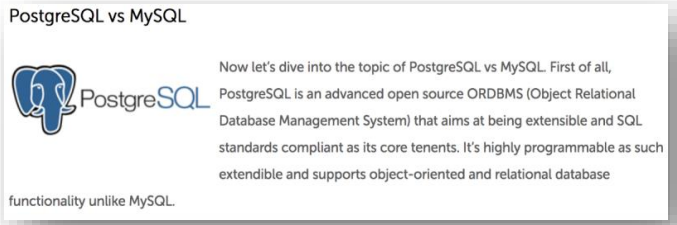
Java IDE
- Eclipse
- IntelliJ IDEA
- NetBeans
- JDeveloper
- …

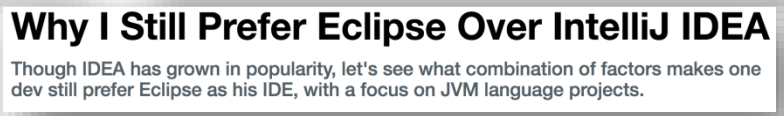
Current Solutions

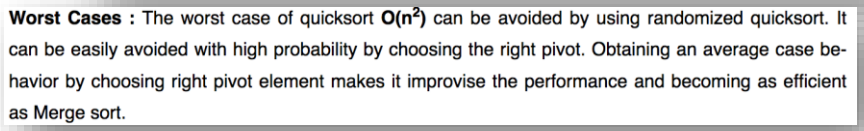
2. Check somebody else's experience – intentional technology comparison

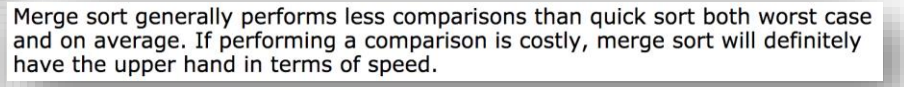
- May not exist
- Fragmented view =>Biased opinions











MONASH University

Inspiration – “Unintentional” Technology Comparison

Any way to compose Solr queries as POSTs in XML,JSON?

1

Yes, you can just switch from GET to POST and it just works.

I wouldn't do it by default however:

- You lose web server logging
- GET is more appropriate than POST for queries because of its safe semantics
- You lose HTTP caching

Software design problem

0

I have an application which will be distributed to a large number of people within my company. I need to have some central data store for this application and dont have the budget for SQL server or anything like this. I noticed that there is a thing called a Local Database in VS2008... will this be suitable for a central data store? the volume of data is not large

1

There is also Postgres, its a bit more robust than MySQL and is free just the same.

System.data.sqlite is a decent option as well for use with Visual Studio, it is also free (but may not be the best option for a large number of clients, although the setup would be easiest by far).

Is UDP Considered to be a “Best-Effort” Service?

1

UDP is generally faster than TCP as it does not have to do the overhead checking of consistency that TCP must deal with. This means that UDP is most often used in programs where transmitting every single last packet correctly is the necessary action. This doesn't mean that UDP is a "best-effort" service, it's something more along the lines of, "You need the information now, and don't care if it's all there"

Hadoop Hortonworks servers down

1

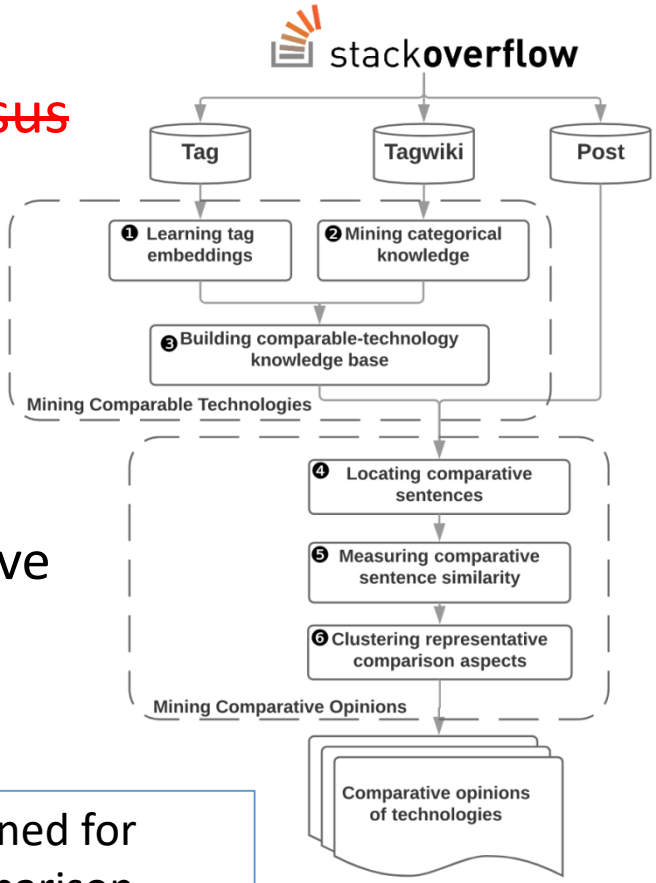
If everything is OK then probably you should check your virtual programs (virtual box or VMWare).

Virtual box is slower than VMWare. When you try to connect Hadoop database and you should see the screen that says you can go http://127.0.... then you can connect otherwise you can see your systems are down.

MONASH University

Approach Overview

- **Mining Comparable Technologies**
  - e.g., nltk versus gate, not ~~nltk versus nlp, nor nltk versus MySQL~~
- **Mining Comparative Opinions**
  - Find comparative sentences, e.g., "GET is more appropriate than POST because of its safe semantics"
  - But comparative sentences ≠ comparative opinions

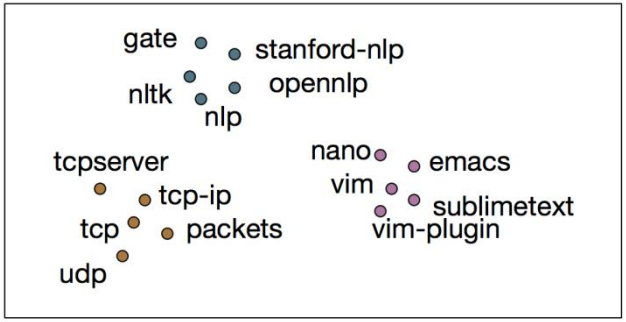


A text summarization technique designed for mining unintentional technology comparison from crowd-scale Q&A discussions

MONASH University

Mining Comparable Technologies

1. **Learning tag embeddings: Use a dense vector to represent each technology**



2. **Mining categorical knowledge: Identify the category of each tag based on Tag Wiki**

Tag Wiki:	Matplotlib	**is**	a	plotting	**library**	for	Python
Part of Speech:	NNP	VBZ	DT	JJ	**NN**	IN	NNP

MONASH University

Mining Comparable Technologies

3. Building comparable-technology knowledge base

- Most close vector
- Same category

Source	Top-5 recommendations from word embedding
nltk	~~nlp~~, opennlp, gate, ~~language-model~~, stanford-nlp
tcp	tcp-ip, ~~network-programming~~, udp, ~~packets~~, ~~tcpserver~~
vim	sublimetext, ~~vim-plugin~~, emacs, nano, gedit
swift	objective-c, ~~cocoa-touch~~, ~~storyboard~~, ~~launch-screen~~
bubble-sort	insertion-sort, selection-sort, mergesort, timsort, heapsort

Mining Comparative Opinions

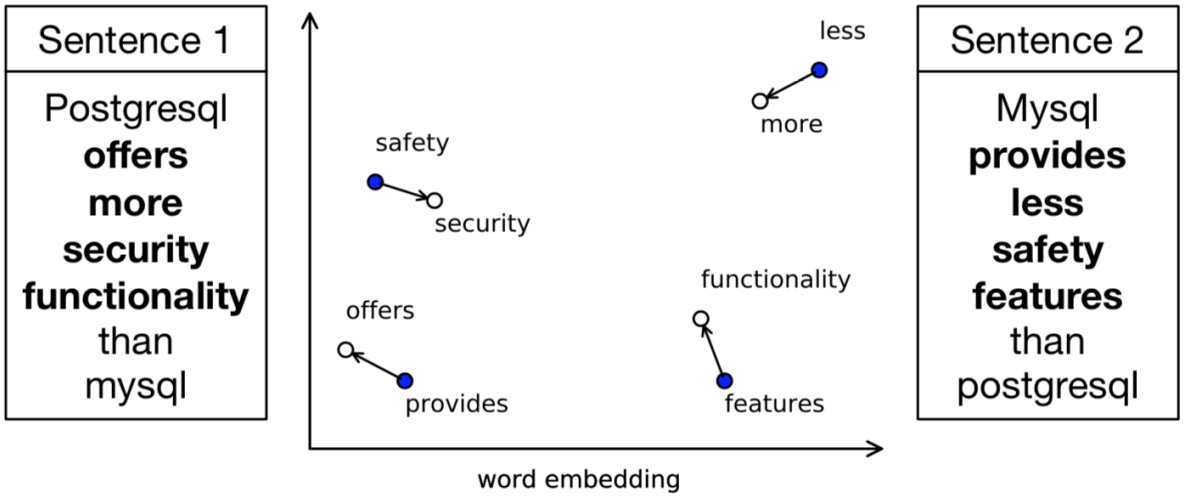
1. Extracting comparative sentences by Part-of-Speech sentence patterns

No.	Pattern	Sequence example	Original sentence
1	*TECH * VBZ * JJR*	innodb has 30 higher	InnoDB has 30% higher performance than MyISAM on average.
2	*TECH * VBZ * RBR*	postgresql is a more	Postgresql is a more correct database implementation while mysql is less compliant.
3	*JJR * CIN * TECH*	faster than coalesce	Isnull is faster than coalesce.
4	*RBR JJ * CIN TECH*	more powerful than velocity	Freemarker is more powerful than velocity.
5	*CV * CIN TECH*	prefer ant over maven	I prefer ant over maven personally.
6	*CV VBG TECH*	recommend using html5lib	I strongly recommend using html5lib instead of beautifulsoup.

MONASH University

Mining Comparative Opinions

2. Measuring sentence similarity by word mover's distance

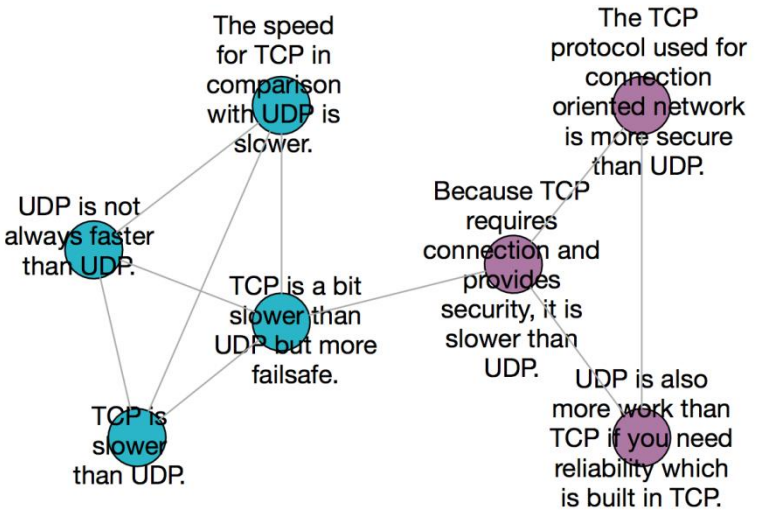


MONASH University

Mining Comparative Opinions

3. Clustering representative comparison aspects and mining cluster topics

- Speed
- Faster
- Slower



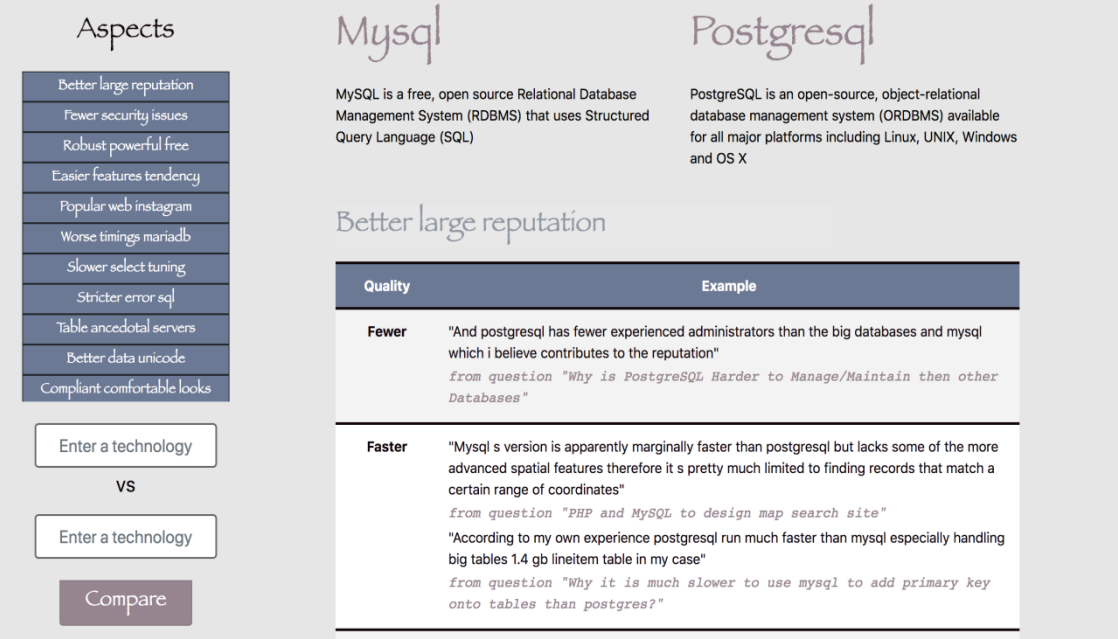
- Secure
- Reliability
- Security

MONASH University

2,074
pairs of comparable technologies

14,552
comparative sentences

Website
https://difftech.herokuapp.com/



Experiments Overview

Quality of each step

- Accuracy of mined comparable technologies
- Accuracy and coverage of mined comparative sentences
- Accuracy of clustering comparative sentences

Usefulness evaluation

- Human-provided intentional technology comparison aspects versus our mined unintentional technology comparison aspects

Experiment

1. Accuracy of Mined Comparable Technologies

- Extraction of tag categories from TagWiki
  - 83.8% accuracy
- Identification of comparable technologies
  - 90.7% versus 29.3% with/without tag category filtering
  - Skip-gram model (90.7%) outperforms continuous bag of words model (88.7%)

Experiment

2. Accuracy of Mined Comparative Sentences

- Examine 50 randomly sampled sentences for each comparative sentence pattern

No.	Pattern	#right	#wrong	Accuracy
1	*TECH * VBZ * JJR*	44	6	88%
2	*TECH * VBZ * RBR*	45	5	90%
3	*JJR * CIN * TECH*	43	7	86%
4	*RBR JJ * CIN TECH*	47	3	94%
5	*CV * CIN TECH*	37	13	74%
6	*CV VBG TECH*	35	15	70%
Total		251	49	83.7%

MONASH University

Experiment

3. Accuracy of Clustering Comparative Sentences

- Word mover's distance can capture the semantic meaning of comparative sentences

- Clustering the graph of similar sentences can explicitly encode the sentence relationships

Method	ARI	NMI	HOM	COM	V-M	FMI
TF-IDF+Kmeans	0.12	0.28	0.29	0.27	0.28	0.41
Doc2vec+Kmeans	-0.01	0.11	0.10	0.14	0.11	0.43
Our model	0.66	0.73	0.75	0.72	0.73	0.79

Usefulness Evaluation

Can our mined comparative aspects answer comparison questions in Stack Overflow?

Question ID	Question title	Tech pair	Tech category	#answers
70402	Why is quicksort better than mergesort?	*quicksort & mergesort*	Algorithm	29
5970383	Difference between TCP and UDP	*tcp & udp*	Protocol	9
630179	Benchmark: VMware vs Virtualbox	*vmware & virtualbox*	IDE	13
408820	What is the difference between Swing and AWT?	*swing & awt*	Library	8
46585	When do you use POST and when do you use GET?	*post & get*	Method	28

Question ID	#Aspects	#Covered	#Unique in our model
70402	6	4 (66.67%)	2
5970383	3	3 (100%)	5
630179	7	4 (57.1%)	1
408820	5	3 (60%)	4
46585	4	4 (100%)	2
Total	25	18 (72%)	14

Our mined “unintentional” comparison aspects have reasonably coverage of human-provided comparison aspects, and sometimes they provide unique aspects not mentioned in intentional technology comparison.

MONASH University

Future Work

- **Improve comparative sentence mining**
  - Technology mentions in separate sentences
  - Co-reference resolution

- **Improve comparison aspect mining and presentation**
  - Preference summarization of comparable technologies

MONASH University