Discovering Keywords Relations

Bc. Peter Kajan peto.kajan@gmail.com
Supervisor: Ing. Michal Barla, PhD

Introduction

Car vs. Geneva Motor Show

Are they related?

Why?

Relatedness of documents, users ...

Search problems: Synonyms, Homonyms, Abstraction Level

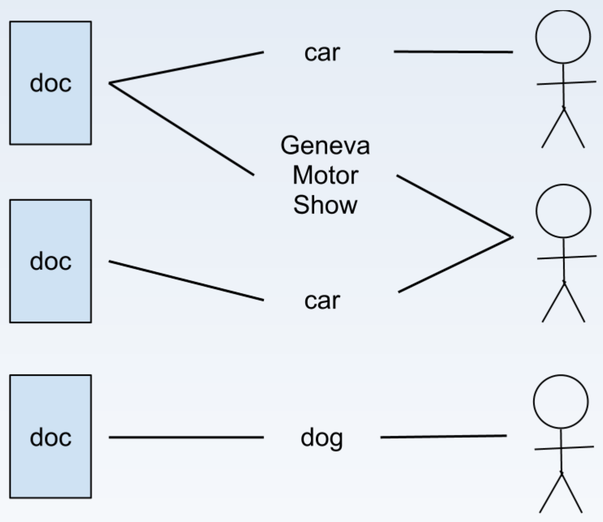
How?

Lexical analyses ✗ Linked Data ✓

Lexicons ✗ Collective intelligence ✓

Tagging vs. Browsing

Tagging



Browsing

Hypothesis:

If significant number of users visit page with keyword A and right after they visit page with keyword B it means that these keywords may describe similar concept.

Tagging vs. Browsing

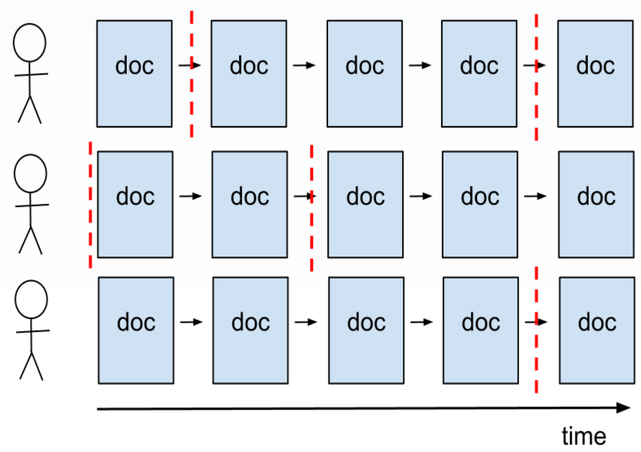
Cheaper data (automatically extracted keywords)

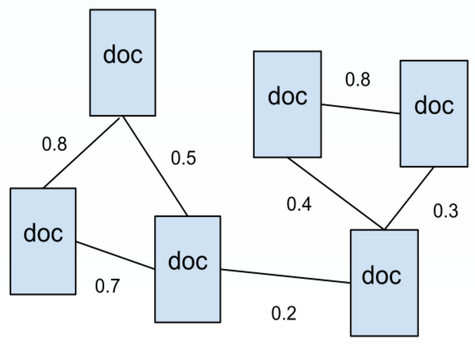
Covering more domain

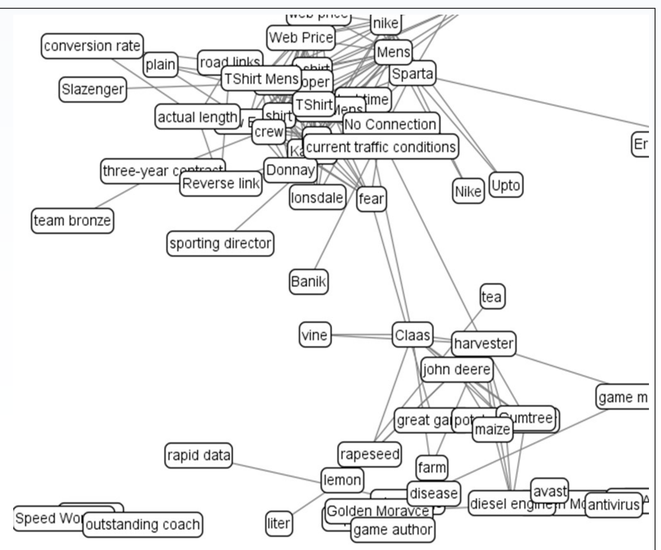
Relevant time ordering

Is time important?

Relatedness

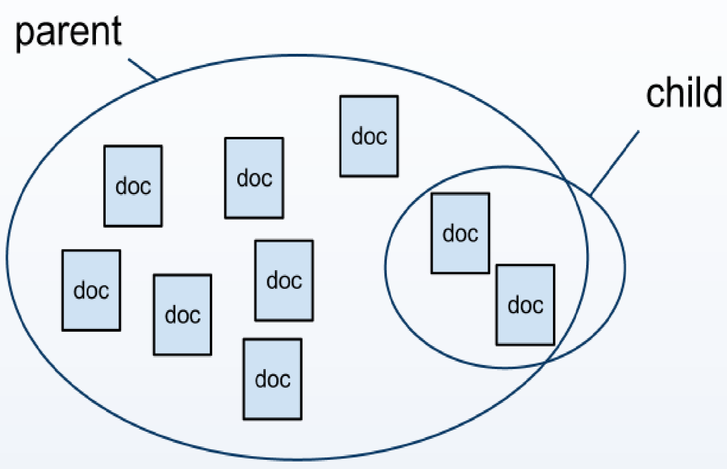






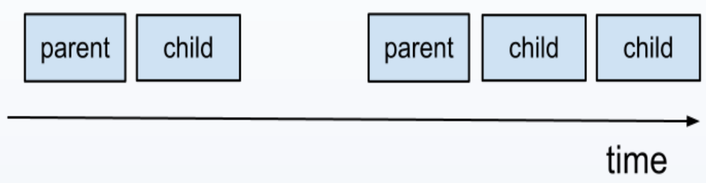
$$s_d(d_a,d_b)=\sum^{\forall(a,b)\in Q_{ab}} f(|a-b|) \qquad c_k(k_a,k_b)=idf(k_a)\cdot\sum^{\forall d_a\in A,\forall d_b\in B} idf(d_a)\cdot s_d(d_a,d_b)$$

Hierarchy



Hypothesis:

Child keyword occurs more likely after parent keyword in browsing sessions.

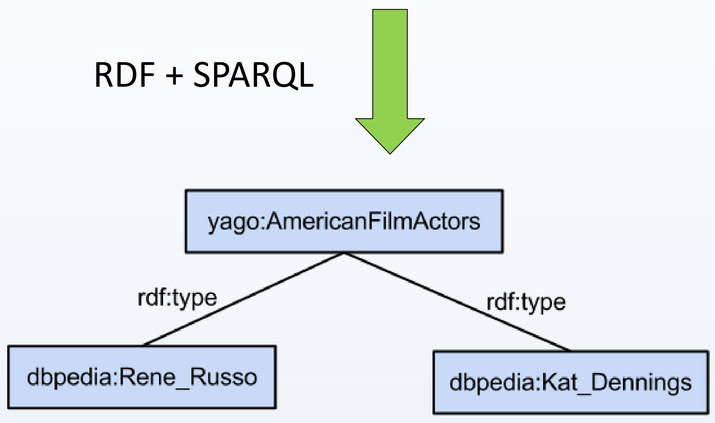


$$p_h(k_a,k_b) = m\cdot\frac{c_k(k_b,k_a)}{c_k(k_a,k_b)} + \frac{\sum^{\forall(a,b)\in R}\frac{1}{a-b}}{|R|}$$

Linked Data

Rene Russo—http://dbpedia.org/resource/Rene_Russo

Kat Dennings— http://dbpedia.org/resource/Kat_Dennings



Results

Experiment

Pewe proxy logs: 560 000 accesses -> 2000 relations only

Context dependent—clustering

Participants picking keywords from related mixed with random

Proposed method (Time)

vs.

Similarity by co-occurrence (No time)

Cluster	First 10 keywords
1	Electrical, faculty, Electrical Engineering, Informatics, Industrial, Control, industrial informatics, Robotics, cybernetics, Technical
2	LG LCD Monitor, Sapphire Radeon HD, intel core 2 quad, Kingston HyperX XMP, Intel P45 Memory, box 2.66 ghz, ii 500w hdd, brand new price, warranty cpu, CPU Cooler
3	Adriana Barraza, Idris Elba, Jaimie Alexander, Chris Hemsworth, Kat Dennings, Natalie Portman, Rene Russo, Anthony Hopkins, Ray Stevenson, Clark Gregg