

Efficient Entity Annotation for Large Scale Web Archives



Elena Demidova, Julian Szymanski, Sergej Zerr and Karol Draszawka

L3S Research Center, Hannover, Germany, Gdansk University of Technology, Faculty of Electronics,
Telecommunications and Informatics, Poland

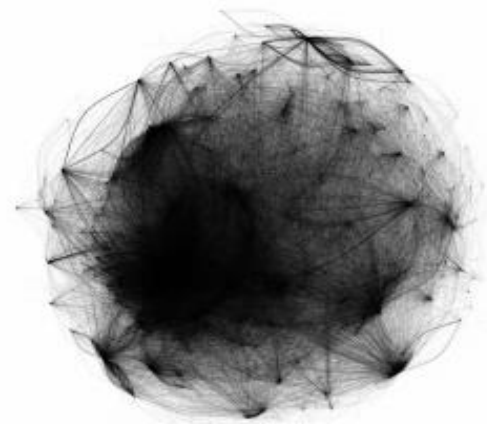
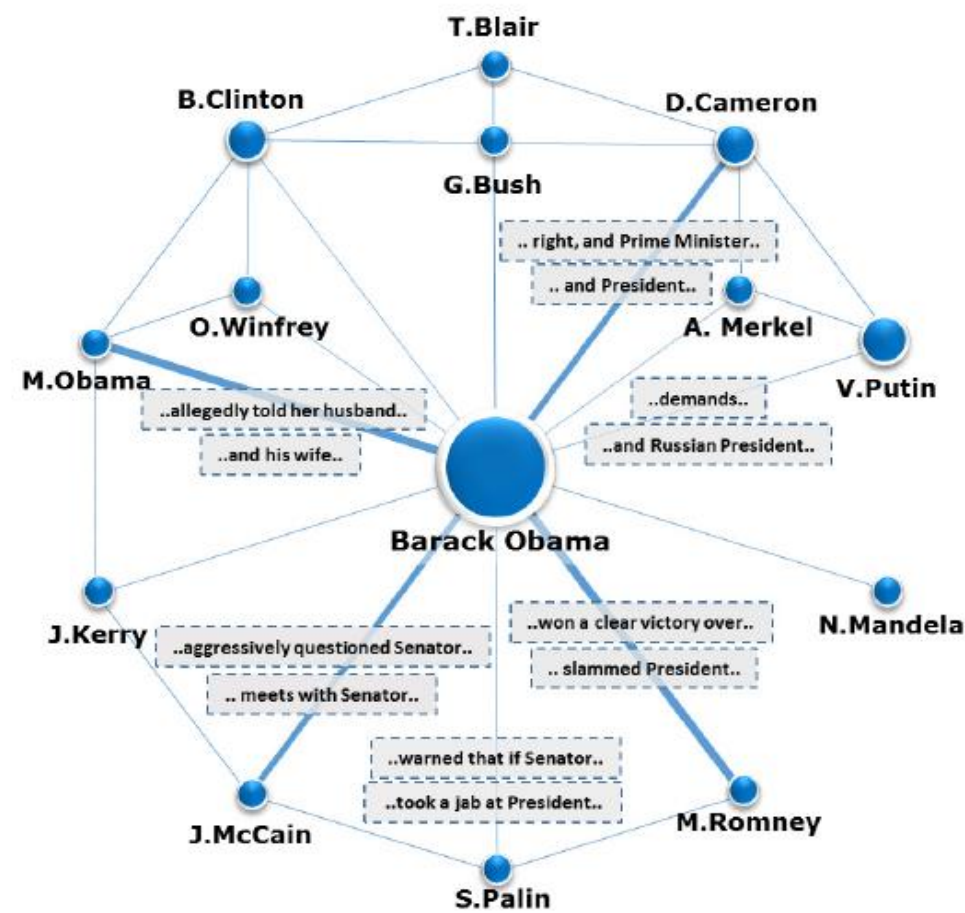
Barack Hussein Obama II (US ⁱ/bəˈrɑːk huːˈseɪn ɵˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

https://en.wikipedia.org/wiki/Barack_Obama

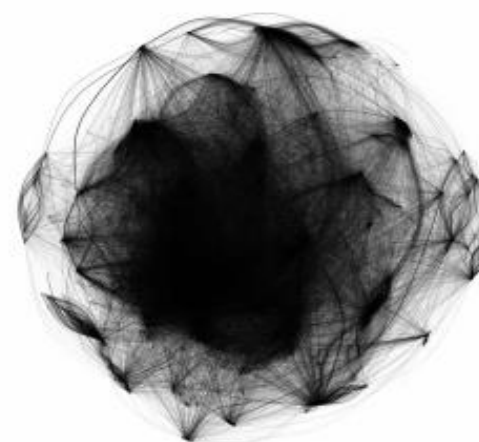
Barack Hussein Obama II (US i/bəˈrɒk huːˈseɪn ɵˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

https://en.wikipedia.org/wiki/Barack_Obama

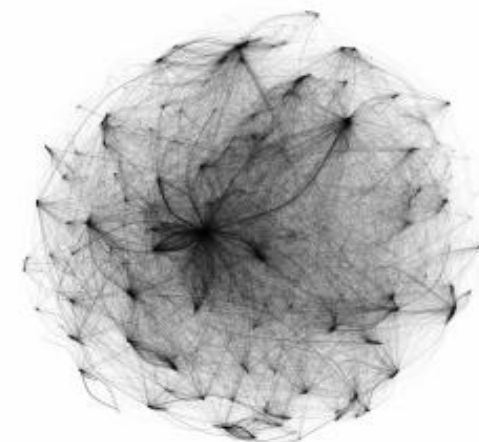
1	Barack Hussein Obama II (US i/ bəˈrɒk huːˈseɪn ɵˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office.
2	Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review.
3	He was a community organizer in Chicago before earning his law degree.
4	He worked as a civil rights attorney and taught constitutional law at University of Chicago Law School from 1992 to 2004.
5	He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.



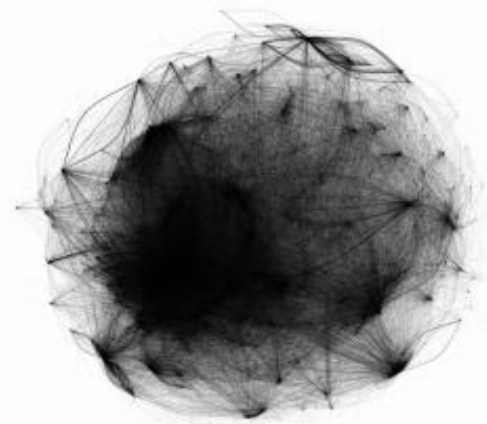
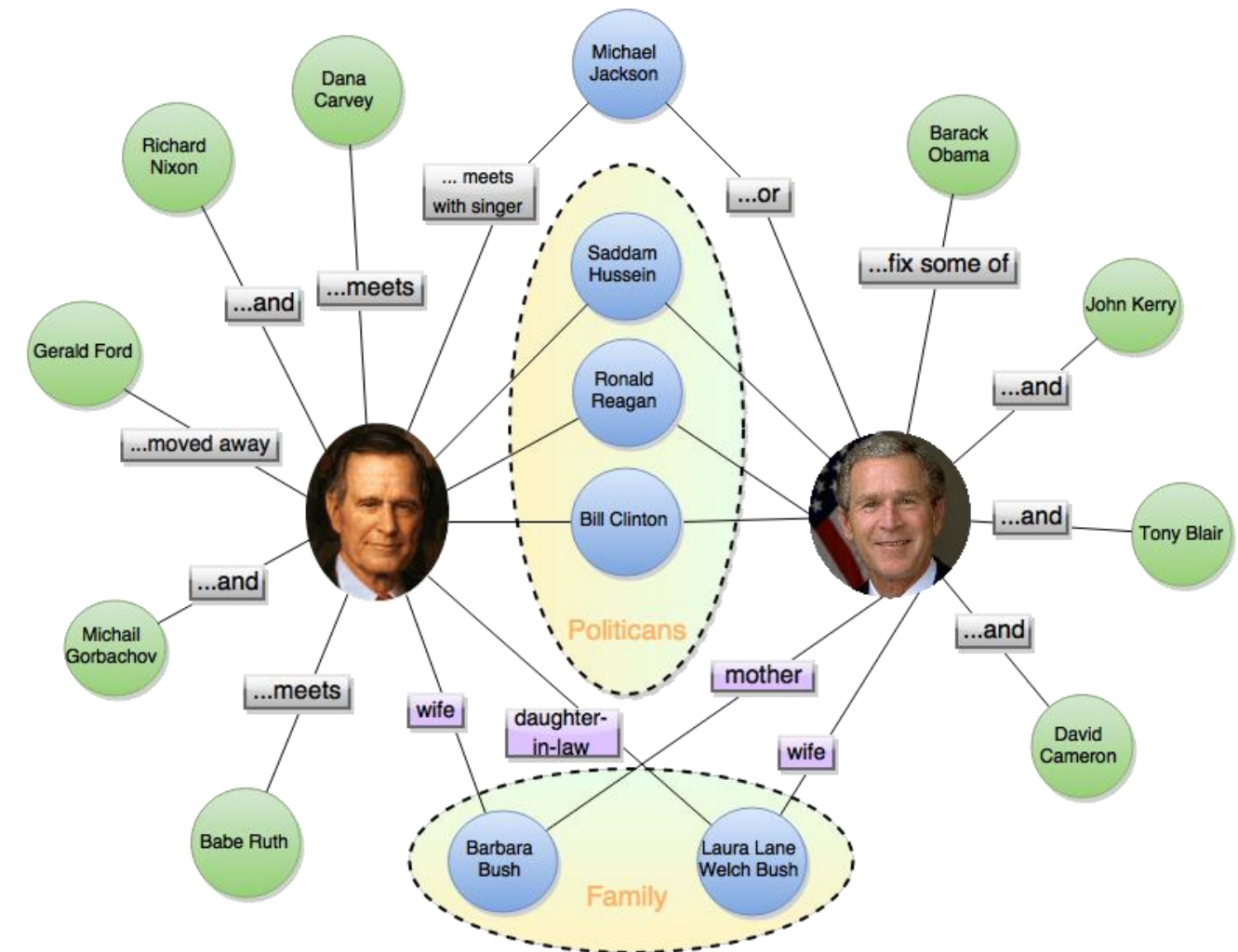
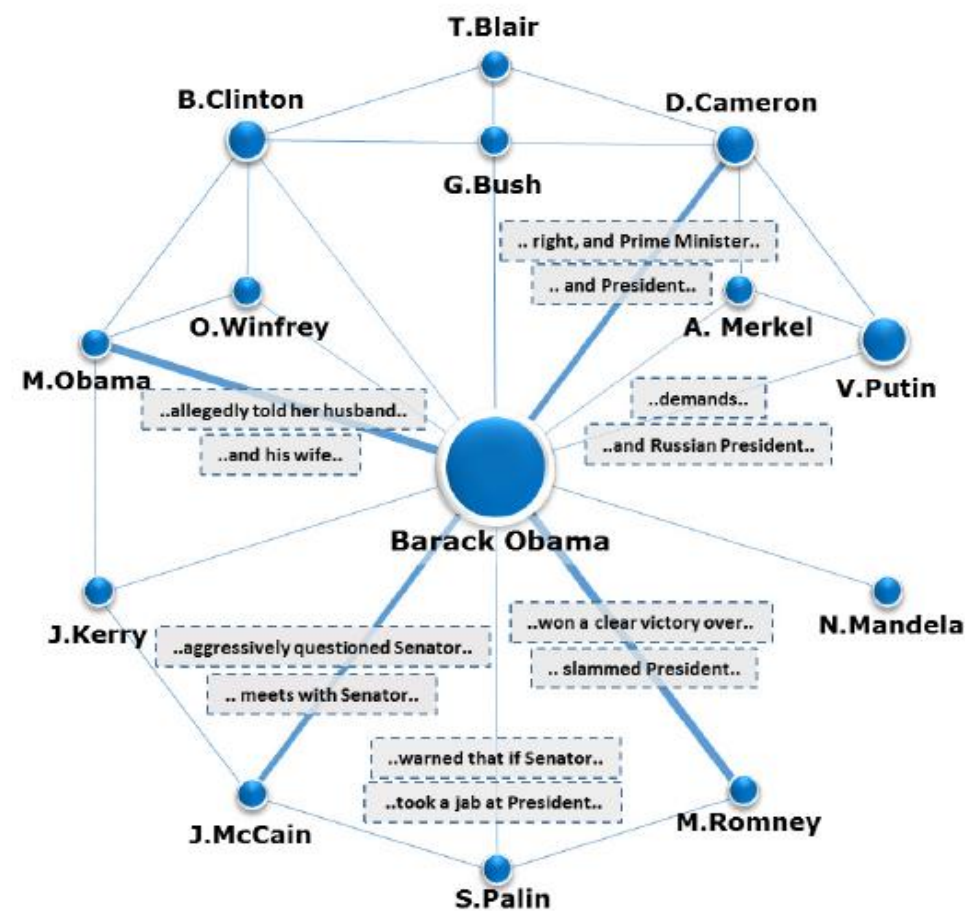
(a) WikiNet BF



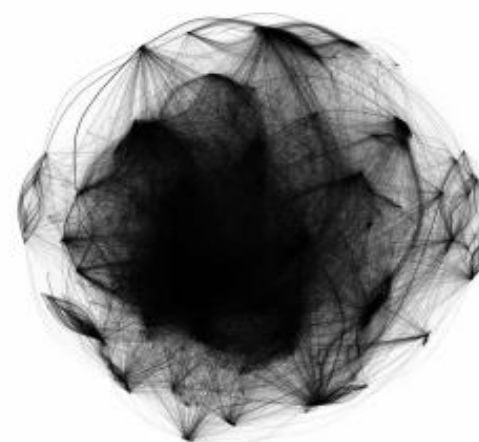
(b) WikiNet $\alpha=0.0$



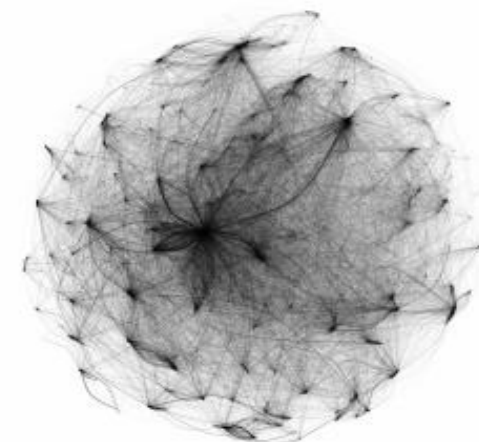
(c) WikiNet $\alpha=0.01$



(a) WikiNet BF



(b) WikiNet $\alpha=0.0$



(c) WikiNet $\alpha=0.01$

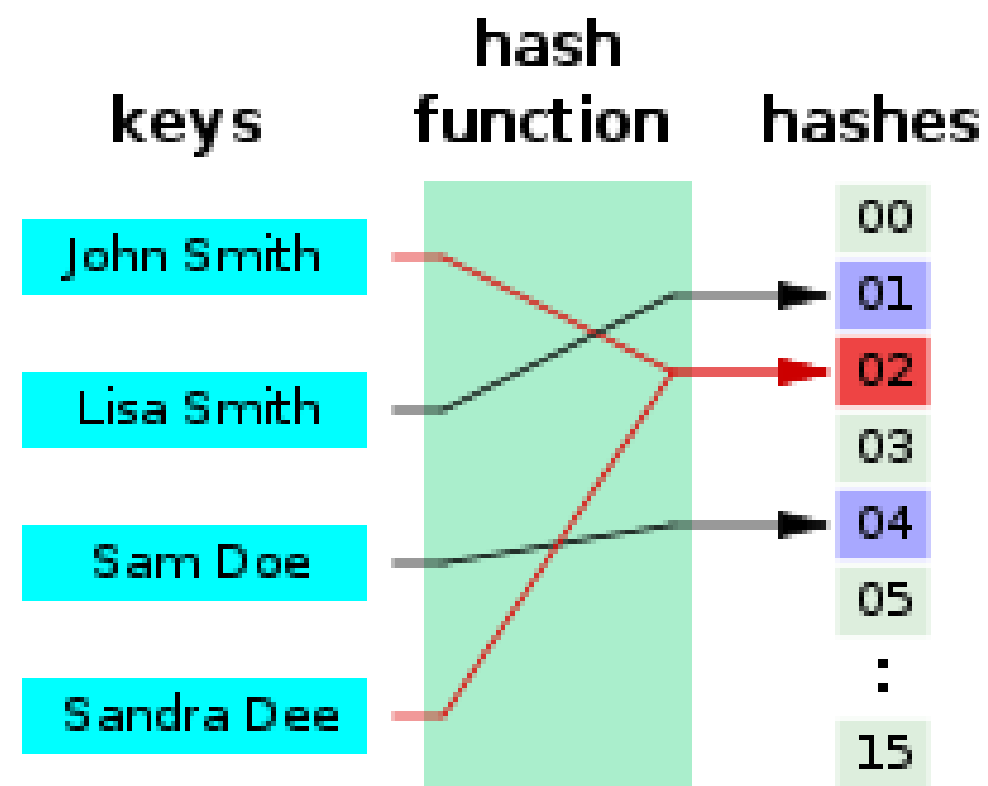
Naive extraction algorithms, running times

Alexandria Project (Foundations for Temporal Retrieval, Exploration and Analytics in Web - ERC 339233)
Access to over 80 TB compressed web pages 1995-2014.

Experiments were conducted on a sample of 2,5 Mio. Web pages.

	Dictionary based	NLP based		
Total runtime	35 357 949 (9h 42min)	133 265 869 (37h 1min)	ms	
Algorithm runtime	3 328 493 (56 min) (x27)	99 232 735 (27h 34min)	ms	
Entities found	3 487 182	8 188 293	persons	
Distinct all entities	156 197	1 448 828 (x10)	persons	
Overlap persons	130 900		persons	

Hashing and Locality Sensitive hashing (LSH)



LSH

00	- ([kay elizabeth, elizabeth kay, elizabeth kempe])
01	- ([tony de sergio, tony di sergio])
02	- ([lawrence bragg, floss hodes])
03	- ([annie hammond, ann hammond])
04	- ([c. williamson, m. williamson, william williamson, g. williamson, h. williamson, jb williamson, ed williamson, williamson jr.])
:	
15	

First experiments

We applied LSH on a set of 291954 person names extracted from the Web archive.

Features:

- 3-gramms (Hammond May= {ham,amm,mmo,mon,ond,may})
- First letters (h,m)
- String count

Examined:

- 291954 different person names
- 40200 different features
- Time ~ 1 minute

Basket size	count
1	267183
2	7749
3	805
4	318
5	161
6	92
7	66
8	44
9	37
10	28
>10	116

a) LSH basket size distribution
(max=120)

Levenstein distance	count
0	1665
1	1426
2	684
3	477
4	499
5	321
6	300
7	274
8	400
9	190
10	207
>10	3240

b) Levenstein distance by collisions
(max=284)

For a sample of 1000 names from baskets of size 1, for around 60% near duplicates (according to levenstein distance) could be found

Example: inga fossa [bossa rosa, hugh foss]: distance:(6.0)

Conclusion

1	Barack Hussein Obama II (US Listen/bəˈrɑːk huːˈseɪn eˈbəmə/; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office.
2	Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review.
3	He was a community organizer in Chicago before earning his law degree.
4	He worked as a civil rights attorney and taught constitutional law at University of Chicago Law School from 1992 to 2004.
5	He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

Observation:

- Dictionary based approach is very fast, however is limited to fixed set of strings
- NLP based approach captures more variations, but is very slow

Idea:

- (1) Extend dictionary based approach with near duplicate matching (LSH) to obtain more entity variations efficiently
- (2) Entity grouping by similarity

Challenges:

- Algorithm parametrization
- Feature selection for similarity measure

Discussions / Questions / Remarks

