Ryadh DAHIMENE
Cédric du Mouza
firstname.lastname@cnam.fr

CEDRIC Laboratory
Teams ISID / VERTIGO

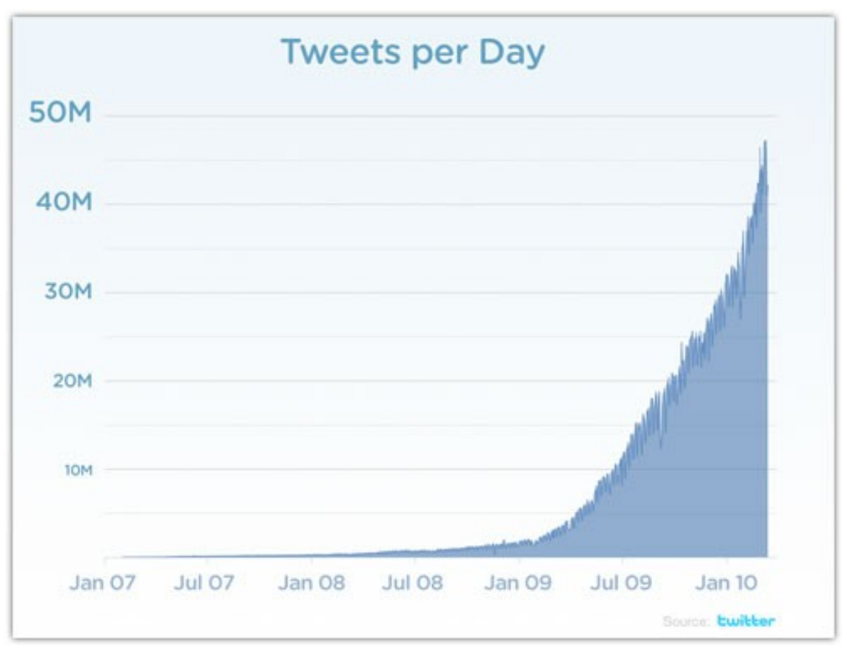
le cnam

MicroFilter: Real Time Filtering For Micro-Blogs

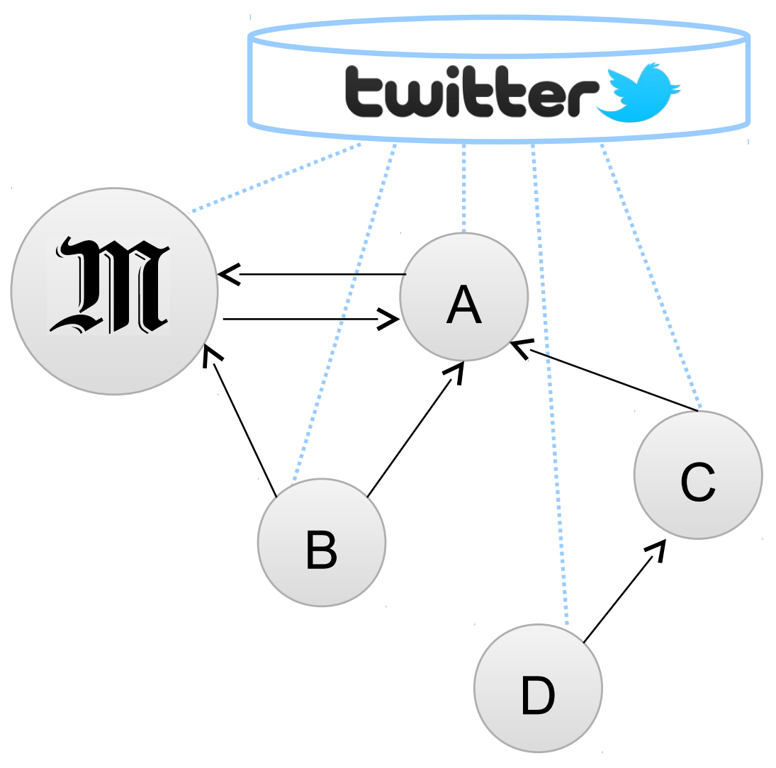
Microblogging systems have become a major trend over the Web. After only 7 years of existence, Twitter for instance claims more than 500 million users with more than 350 billion delivered update each day. As a consequence the user must today manage possibly extremely large feeds, resulting in poor data readability and loss of valuable information and the system must face a huge network load.
In this demonstration, we present and illustrate the features of MicroFilter, an inverted list-based filtering engine that nicely extends existing centralized microblogging systems by adding a real-time filtering feature.

Microblogging Systems specific aspects:

- **Short messages:** less than 140 characters (average length of 14,7 words).
- **Directed graph of users:** a user A follows another user B implies that A receives all of B updates.
- **Heterogeneity in accounts sizes:** in term of update frequency and number of followers.
- **Graph evolution:** users change accounts they follow often.
- **Centralized state:** on *Twitter*, every new post is handled by *Twitter* servers and than transmitted to the followers.



Tweets per day – Jan07 to Jan10

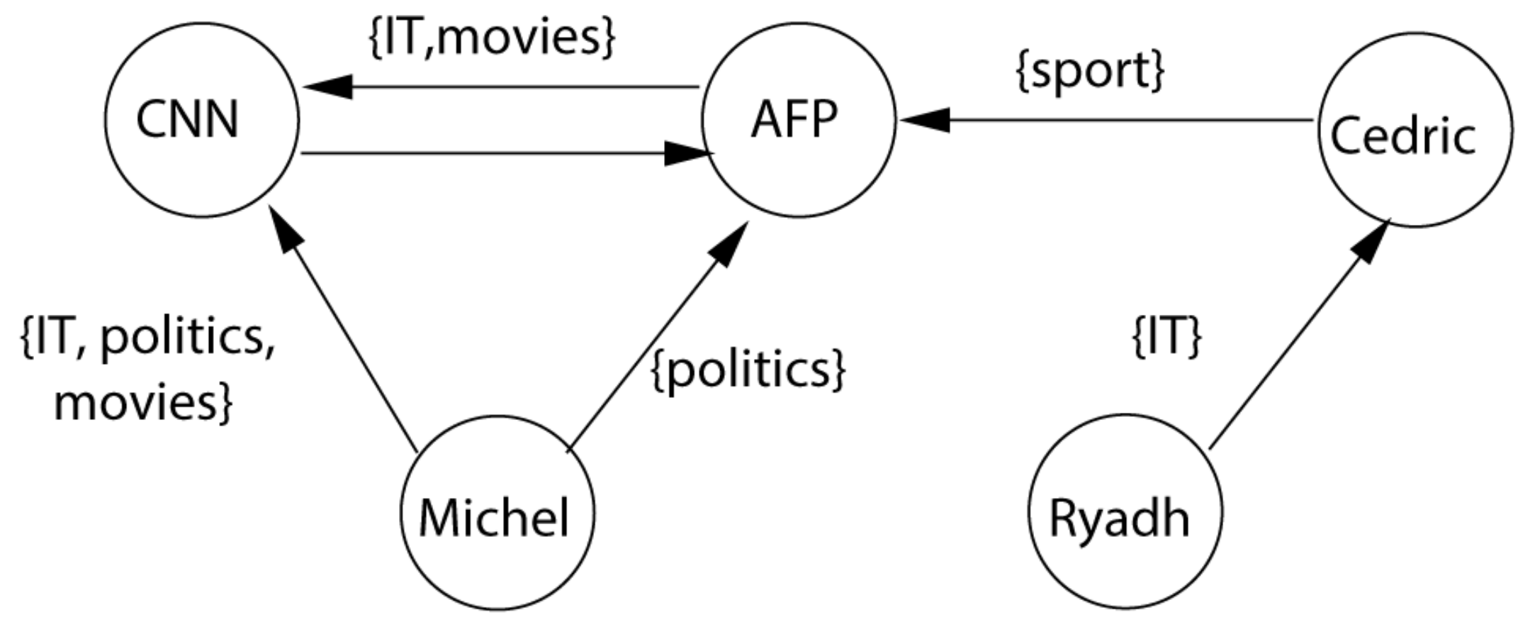


Example of a *Twitter* directed graph

Our approach: Filter an account *tweets* in order to receive only a subset of emitted tweets, based on interest queries (keywords).

Our goal:

- Improve the user experience and reduce the load of the network.
- Use inverted list indexes to store queries related to a graph arc. We proposed three different structures (*PFT*, *PTF*, *TPF*).

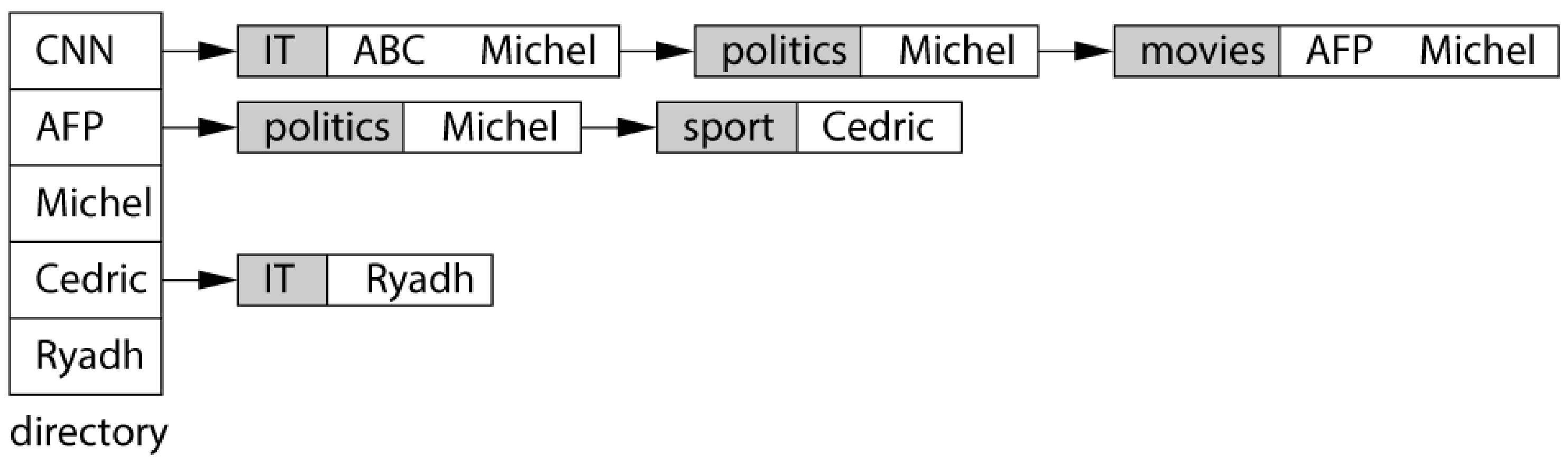


An Example of a filtered social graph

Element	♯
Users	2,170,784
Tweets	15,717,449
Graph arcs	148,508,857

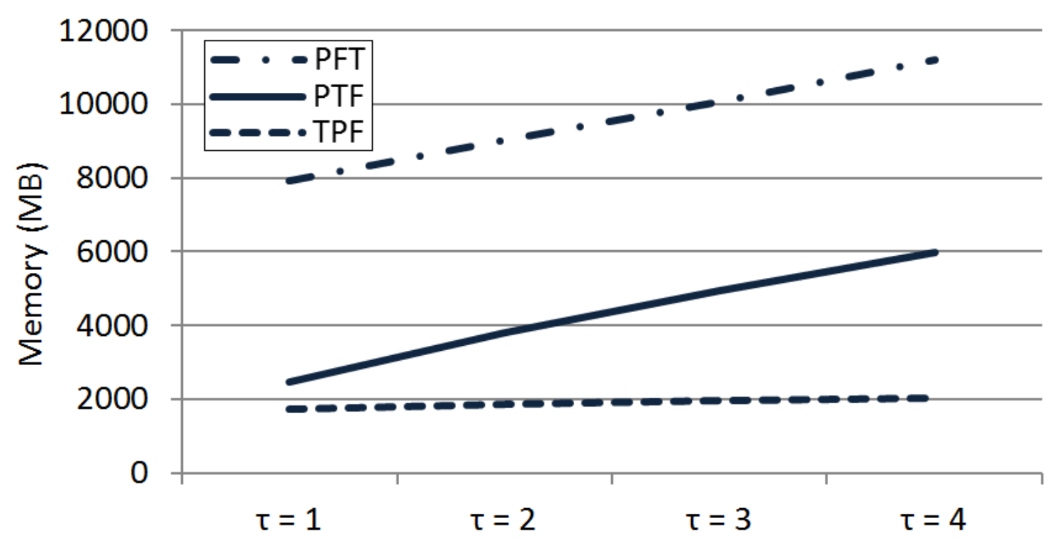
User	Follower	Queries
12	36255503	bigday twitter
12	36255965	conference deadline download
12	36256156	DB conference
12	36256607	software twitter

Dataset sample

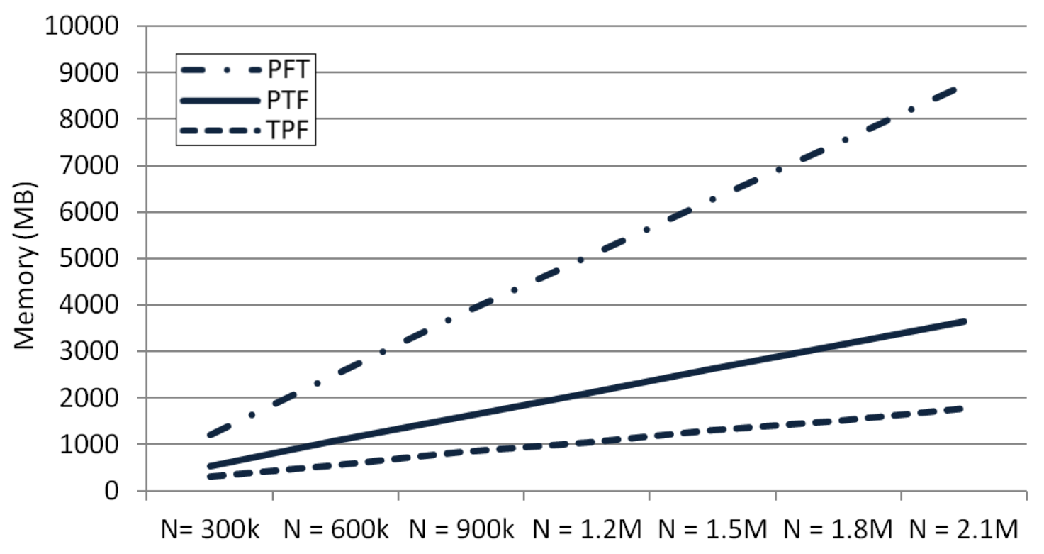


An example of the *PTF*-Index

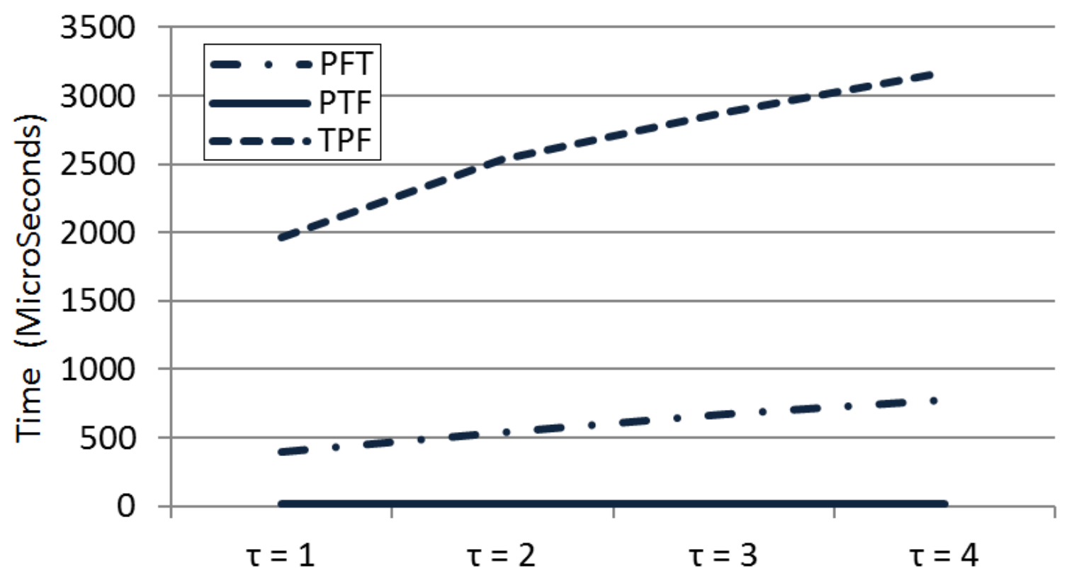
Experiments:



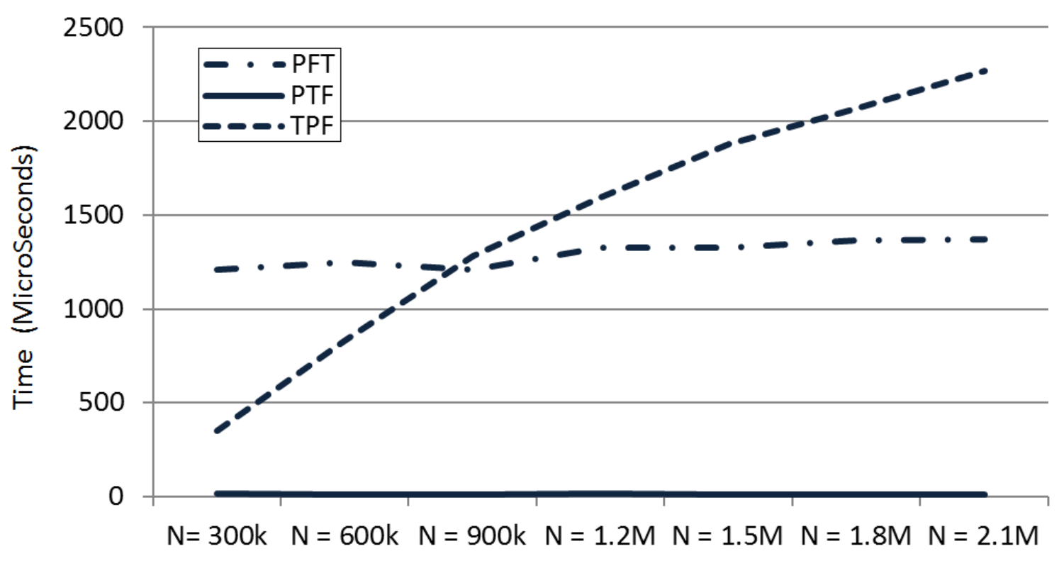
Occupied memory w.r.t. τ (filter size)



Occupied memory w.r.t. N (number of accounts)



Matching time w.r.t. τ



Matching time w.r.t. N

We compared three inverted lists-based structures and validate them with real and synthetic datasets. PTF-index appears to achieve the best scalability, despite memory requirements and twice more important than TPF-index, it outperformed with two orders of magnitude other proposals for matching time.

Future Work:

We intend in future work to consider other optimizations like clustering or summarization which group different filters inside a posting list to achieve better performance.
Adding conjunction and negation in filter expressions is another future challenge. We are also working on real time content recommendation for microblogging systems on top of our structures.

Related publication:

[DMS12] Ryadh Dahimene, Cédric du Mouza, Michel Scholl. Efficient Filtering in Micro-blogging Systems: We Won't Get Flooded Again. Intl. IEEE Conf. on Scientific and Statistical Databases (SSDBM'12), 2012, pp.168-176