

Enabling Webscale Research in Europe

Julien Masanès

European Archive Foundation

julien@europarchive.org

FI PPP, Nice, 12/3/2010

Webscale data

- The web represents a unique source of access to media content of all sorts, that a growing number of scientific communities, public agencies and industries need to mine at large scale.
- The ability to acquire, process and mine large scale data from the web is becoming a strategic advantage in many domains from business intelligence to epidemiological tracking and monitoring.

Who can do research on Webscale data?

- Webscale is already proving to be a challenge for many research group as the infrastructure, the cost and the skills required represent a significant barrier to entry.
- But when it comes to doing this using time series, all but a few (mainly large search engines) can do it at all.
- In other words, only large search engines (none being European) are able to do research at this scale, hence comforting their advance by developing and testing new algorithms for search, ranking, mining etc.

Research engine for all application domains

- Content: monitor and analyze the structure and evolution of networked media
- The web reflects all aspect of society, it can be use to analyze:
 - economical trends,
 - social use of technology
 - emergence of new research of new research fields
 - tracking of reputation on the web
 - etc.

Health

- Health: epidemiological tracking
 - analyze emergence and diffusion of diseases based on the conversational web
 - track how new drugs are perceived, potential secondary effects, illegal market etc.

Our contribution to FI PPP

- European Archive Foundation has expertise in:
 - Large scale archiving crawling (for instance crawl for the UK government, German TV, European domains)
 - Distributed analytics (LAWA FP7 project)
 - Management of none technical issues related with content (privacy, IP rights etc)
- Infrastructure in Amsterdam and Paris (1PB)

Thanks

Julien Masanès

European Archive Foundation

julien@europarchive.org

FI PPP, Nice, 12/3/2010