Selective Information Dissemination on the Real-time Web - A Database Perspective

Bernd AMANN, UPMC – LIP6
Keynote WEBIST 2015
Lisbon, Portugal
May 20, 2015
LIP6 Database Research Group

Web Information Streams
  continuous top-k queries
  multi-query optimization
Web Crawling and Archiving
  crawling dynamic contents
  archive construction and querying
Cloud Data Processing
  large-scale transactions
  data distribution and replication
Data provenance and quality
  provenance generation and
  quality of data-centric workflows
Text document representation
  computational linguistics
  machine learning
My personal « Web » timeline

Hypertext Query Languages (PHD)
XML repositories and Active views
Semantic XML data integration
Intensional XML data
Web Service Ranking
RSS data acquisition and aggregation
XML Workflow provenance
Continuous Top-k Query Processing

1990
1995
2000
2005
2010
now

HTML
XML&RDF
Web Service
RDF
Social Web
Real-time web
The Web Revolution

Web of Contents
publish, search and explore contents

Web of Services
access and update data
interact with data

Web of People
publish and share
interact with people

Web of Things
interact with things
"Interactive" Web

Contents
Services
People
Things

explore, search
reserve, buy
connect, share, comment, recommend
connect, survey, control

Google
Le Monde.fr
Wikipedia

Amazon
ACCOR

Twitter
Facebook

YouTube
flickr
del.icio.us

Contents Services People Things

WEBIST Keynote
May 20, 2015

Bernd AMANN - Selective Information Dissemination on the Real-Time Web
Interaction = dynamic data

explore, search
reserve, buy
connect, share, comment, recommend
connect, survey, control

static → dynamic → real-time
Real-time web

Static web
- Large, complex documents
- Low information decay
- Ad-hoc search
- On demand processing

Real-time web
- Simple, short messages
- High information decay
- Publish-subscribe dissemination
- Continuous processing
Twitter: real-time information dissemination

<table>
<thead>
<tr>
<th></th>
<th>followers</th>
<th>following</th>
<th>tweets/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>@HuffingtonPost</td>
<td>5 700 000</td>
<td>5 500</td>
<td>200-500</td>
</tr>
<tr>
<td>@justinbieber</td>
<td>53 000 000</td>
<td>210 000</td>
<td>5-15</td>
</tr>
<tr>
<td>@BarackObama</td>
<td>69 000 000</td>
<td>642 000</td>
<td>5-20</td>
</tr>
<tr>
<td>@NBA</td>
<td>14 000 000</td>
<td>1 300</td>
<td>130-230</td>
</tr>
<tr>
<td>@MTV</td>
<td>12 000 000</td>
<td>31 000</td>
<td>100-250</td>
</tr>
<tr>
<td>@CNN</td>
<td>17 000 000</td>
<td>1 000</td>
<td>40-70</td>
</tr>
<tr>
<td>@InternetRadio</td>
<td>1 300</td>
<td>32</td>
<td>800 – 1000</td>
</tr>
<tr>
<td>@AmexOffers</td>
<td>56 000</td>
<td>0</td>
<td>100 – 40 000</td>
</tr>
</tbody>
</table>

http://twittercounter.com/

500 000 000 tweets/day
Rest of this talk

Online information aggregation, ranking and filtering

News
Web syndication (RSS)
Micro-blogging (Twitter)
Social media

Outline:

RoSeS : Content-based RSS aggregation
Meows : Continuous top-k queries over the real-time web

Perspectives
Really Open, Simple and Efficient Syndication

Jordi Creus\textsuperscript{1}, Roxana Horincar\textsuperscript{1}
Bernd Amann\textsuperscript{1}, Nicolas Travers\textsuperscript{2}, Dan Vodislav\textsuperscript{3}

\textsuperscript{1} LIP6 – Université Pierre & Marie Curie
\textsuperscript{2} CEDRIC – Conservatoire National des Arts et Métiers
\textsuperscript{3} ETIS – Université de Cergy-Pontoise
Content-based RSS Aggregation
RSS Aggregation

RSS Aggregator

subscribe
publish

RSS
Aggregator

Le Monde.fr
YouTube
flickr
The New York Times
EURO SPORT
CNN
RSS web syndication

Really Simple Syndication
standard web feed format to publish frequently updated information

RSS feed
XML document including full or summarized text and metadata with links

RSS aggregators
Feedly: 40 M feeds / 15 M users
Google Reader (< June 2013)
Yahoo Pipes
https://pipes.yahoo.com/

RSS usage:
top 10k sites: 21 %
top 100k sites: 22 % top
1M sites: 30 %
entire Internet: 6 % (20M)

Site types:
Business
News
Social
Technology

source http://trends.builtwith.com
RSS Feeds

http://www.wikicfp.com/cfp/rss?cat=web

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>CFPs on Web : WikiCFP</title>
    <link>http://www.wikicfp.com/cfp/call?conference=web</link>
    <description>A Wiki to Organize and Share Calls For Papers</description>
    <item>
      <title>SaW 2015 : International Workshop on Semantic and Web</title>
      <link>http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=46102</link>
      <description>International Workshop on Semantic and Web [Porto - Portugal]
      </description>
      <guid isPermaLink="false">cfp-528064-S@wikicfp.com</guid>
    </item>
    ....
  </channel>
</rss>
```
Really Open, Simple and Efficient Syndication

[WISE'2010]  
[DEXA'2011]  
[CIKM'2011 demo]  
[ICWE'2012]  
[WWWJ'2014]
Seven more terrorism suspects have been arrested and detained in the United Kingdom in what is now the fourth security-related incident this week as the world counts down to the London Olympics, which begin three weeks from today.

Unemployment Unchanged at 8.2 Pct

The Labor Department announced unemployment figures for the month of June.

Obama Bus Rolls Across Ohio With Seal

It's been dubbed "Ground Force One" and now has the official insignia to match its partner in the air. The jet-black armored motor coach ferrying President Obama from Air Force One in Toledo, across northern Ohio and into Pennsylvania, set an upgrade for its first
Requirements
- Query collections of feeds
- Apply text filters
- Annotate items with complementary information

Publication language:
- union
- selection
- join & window

Example
create feed newsOfSyria
from nytimes | cnn | telegraph
where title contains “syria” or title contains “assad”
create feed myMovies from allocine as $a$
join last 3 weeks on myFriendsTweets with $a[title \text{ similar} \(Utils:window.title)]$
where $a[description \text{ not contains} \"julia roberts\"]$
System Overview

- Crawler
  - Items to evaluate
  - Evaluation & Optimization
- Query Processor
  - Items to publish
- Dissemination
- Catalog Interface
  - Publication queries
  - Subscriptions

WEBIST Keynote
May 20, 2015
Bernd AMANN - Selective Information Dissemination on the Real-Time Web
Query Processor

WEBIST Keynote
May 20, 2015
Bernd AMANN - Selective Information Dissemination on the Real-Time Web
Multi-Query Optimization

src1 -> U -> σ(pred1) -> U -> σ(pred3) -> pub1

src2 -> U -> σ(pred2) -> U

src3 -> U -> σ(pred2 ∧ pred4)

src4

src5

WEBIST Keynote
May 20, 2015
Bernd AMANN - Selective Information Dissemination on the Real-Time Web
Build Shared Filter Plans

Rewriting rules

Distribute selection over unions
Flatten cascading selections
View decomposition
Commute join and selection, …
Extended Filter Plans
(with cost estimation)

rate(src) \cdot \text{selectivity}(b \land c) = 1 \cdot \text{selectivity}(b) \cdot \text{selectivity}(c)
Find a Steiner tree

Minimum tree spanning a given subset of vertices called terminal vertices (query filters)
Extended Filter Plan Optimization

Tree cost = 2.23
Extended Filter Plan Optimization

Tree cost = 2.12
VCA Algorithm

Steiner tree problem is NP-complete

STA Algorithm

greedy state-of-the-art algorithm with approximation guarantees

VCA – Algorithm

expansion/reduction step at each iteration

benefit(x, y) = (n – 1) * selectivity(y) – n * selectivity(x)

n : number of sons of y also subsumed by x
Experiments: number of queries
RSS Crawler - Challenges

Maximize aggregation quality
  - stream completeness (long-term information loss)
  - window freshness (short-term outdated information)

Limited resources
  - bandwidth
  - storage
  - memory or computing capacity

Highly dynamic content
  - publication model
  - change estimation

Refresh strategies
Crawler architecture

Refresh strategy
uses source publication models in order to take refresh decisions

Online change estimator
updates the source publication models based on the observations of the real source publication behaviors
Crawler architecture

Refresh strategy
uses source publication models in order to take refresh decisions

Online change estimator
updates the source publication models based on the observations of the real source publication behaviors
"Best effort" Refresh strategy

"Best effort" strategy (Lagrange Multipliers):

Let $U_t(s, q, t, T_r)$ be a monotonic utility function and $\tau$ a positive constant. At each time instant $t \geq T_r$, refresh all sources $s$ where $U_t(s, q, t, T_r) \geq \tau$.

→ obtains maximum quality compared with all other strategies that use an equal cost (number of refreshes).

Applications

- web pages, cache synchronization, RSS (retrieval delay)
Saturation & Monotonicity

Saturation

feed divergence $\geq$ publication window size

Divergence

stream – monotonic

window – non monotonic after saturation
Refresh strategy for saturated and non saturated sources

Saturated sources: top-k divergence
refresh sources with maximal divergence score

Non saturated sources: best effort
refresh sources with $Uti(s, q, t, T_r) \geq \tau$

→ obtains maximum utility compared with all other strategies that use an equal cost (average number of refreshes).
Crawler architecture

Refresh strategy
uses source publication models in order to take refresh decisions

Online change estimator
updates the source publication models based on the observations of the real source publication behaviors
Source publication activity analysis

RSS source feeds:
- 1658 random feeds over 4 weeks
- 963 feeds over 4 weeks: online French and international newspapers

Shape discovery algorithm
Publication shapes:

- Peaks
- Uniform
- Waves
Experiments
“2 steps” Refresh strategy with Online estimation

![Graphs showing feed completeness and window freshness](image)

**Uniform**
- Feed completeness
- Window freshness

**Waves**
- Feed completeness
- Window freshness

Average refresh frequency (hours)
Meows: Continuous Top-k Filtering over the Real-time Web

Nelly Vouzoukidou\textsuperscript{1}, Bernd Amann\textsuperscript{1}, Vassilis Christophides\textsuperscript{2},

\textsuperscript{1} LIP6 – Université Pierre & Marie Curie
\textsuperscript{2} University of Crete, Heraklion, Greece

[CIKM'2012]
[CIKM'2014 demo]
Selective Information Dissemination on the Real-Time Web

MeowsReader > All my Meows > Malaysia Airlines

Another Dead End, But Malaysia Airlines Searchers Not Giving Up
The first objects fished out of the Indian Ocean that searchers hoped would give them a clue about the missing Malaysia Airlines jet turned out to be another dead end, but organizers of the international effort said they were "well, well short"

ABC News | Mon, 31 March 2013, 2:20PM

MH370: Hopes dashed as orange objects turn out to be fishing equipment
Kuala Lumpur, Malaysia (CNN) -- Potential leads on the missing Malaysian jetliner keep coming. So do the setbacks and frustrations. Monday's search ended without finding anything significant, Australian officials said...

CNN | Mon, 31 March 2013, 2:10PM

Race on to find missing plane’s black box before it stops pinging
Royal Australian Air Force crew members read navigation maps during Thursday’s search effort. Photo: Pool/Getty Images. MORE ON: Flight MH370 · Black box detecting

Trends:
- Health care
- Empire State Building
- Crimea
- "The Walking Dead" Russia
- Pervez Musharraf
- Oculus
- Flight 370
- FedEx Lawsuit
- Ice Storm Study
- Software
- Patents
Continuous Top-k Query Processing

Context:
Continuous information retrieval in text streams (news, tweets)
Top-k queries: produce top-k result wrt. some ranking function

Contribution:
Formal top-k continuous query filtering model
Inhomogeneous ranking functions combining query-dependent (similarity) and query-independent (importance) item scores
Efficient index structures for continuous top-k text queries with inhomogeneous ranking functions
Inhomogeneous ranking function

Information stream
articles/items

Item score - \( S_{item}(i) \)

- source
- authority

- diversity
- other information

Query score --
\( S_{query}(q,i) \)

Query Filtering

\[ S_{total}(q, i) = \alpha \cdot S_{item}(i) + \beta \cdot S_{query}(q, i) \]
State of the Art:
COL-Filter Using Threshold Algorithm

State of art system (for homogenous score functions):
COL-Filter [HMA2010]:
TA-like algorithm [Fagin01]

\[
\begin{align*}
q_1 &= \{t_1, t_2, t_3, t_4\} \\
q_2 &= \{t_1, t_2, t_4\} \\
q_3 &= \{t_2, t_3\} \\
q_4 &= \{t_2\} \\
q_5 &= \{t_1, t_3\}
\end{align*}
\]

\[S_{total}(q, i) > S_{\text{min}}(q)\]
\[S_{total}(q, i) = \sum_{w_{q,t}} w_{i,t}\]
Term Weights/Query Score Constraints

Filtering constraints depending on minimal query score and term weights
Query Indexes

Spatial indexing problem
We consider variations of:

- Rectangular grids
- R-Trees

Different space, filtering and update time costs
Experiments: Time/Space trade-off

Average per item query detection time over number of stored queries

Using 20% more memory in SortQuer, we achieve 55% better performance than state-of-art COL-Filter
Experiments: Query Scale

Average per item query detection time over number of stored queries

![Graph showing query scale experiments]
Approximate Query Filtering

$w_{q,t}$

$L^B$

$HTA$

$MQW$

$S_{min}(q)$

Always updated

Never updated
Inhomogeneous **dynamic ranking function**

Information stream articles/items

**Item score** - $S_{item}(i)$

- source authority
- diversity
- other information

**Dynamic Item scoring function**

**Query score** - $S_{query}(q, i)$

**Query Filtering**

decay

$$S_{total}(q, i) = \alpha \cdot S_{item}(i) + \beta \cdot S_{query}(q, i) + \gamma S_{feedback}(i)$$
Perspectives and Conclusion
Web of Things

Different Things Need To Be Protected

Rapid adoption rate of digital infrastructure
5 x faster than electricity & telephony


© CISCO
What's next: "Ubiquitous" Web

Data and service Ubiquity
universal connectivity
communication
pervasive computing
data cloud

Real-time/real-world interaction
"smart" car/home/city/industry
personal/ambient/social awareness

Data and Service Explosion

© W3C
Data Dissemination in the "Ubiquitous" Web

Data complexity
- heterogeneity (contents, structure, semantics)
- metadata: source, annotation
- user context: social, spatial, …

Data quality
- correctness
- completeness
- timeliness

Data control
- privacy
- security

High-dimensional ranking and recommendation
- contents, space, time, semantics, feedback, …

Data & stream processing
- new hybrid architectures
- new algorithms

Provenance
- why/where/how provenance

Logics and semantics
- reasoning and verification

User interfaces
- interaction
- publication & subscription administration
Thank you for your attention!
Selective Information Dissemination on the Real-Time Web