A Software Architecture for Progressive Scanning of On-line Communities

Roberto Baldoni, Fabrizio d'Amore, Massimo Mecella, *Daniele Ucci*
Sapienza Università di Roma, Italy

CIS SAPIENZA

RESEARCH CENTER FOR CYBER INTELLIGENCE AND INFORMATION SECURITY

Motivations

On-line communities

- a fundamental source of information in business and information security intelligence
- contain information that can be used for inferring trends and evolution about specific topics
- A social community, and what it publishes, often influences other communities, and vice-versa …
- … thus creating a network of causal relationships that can contain useful information about the evolution of a specific phenomenon

Model

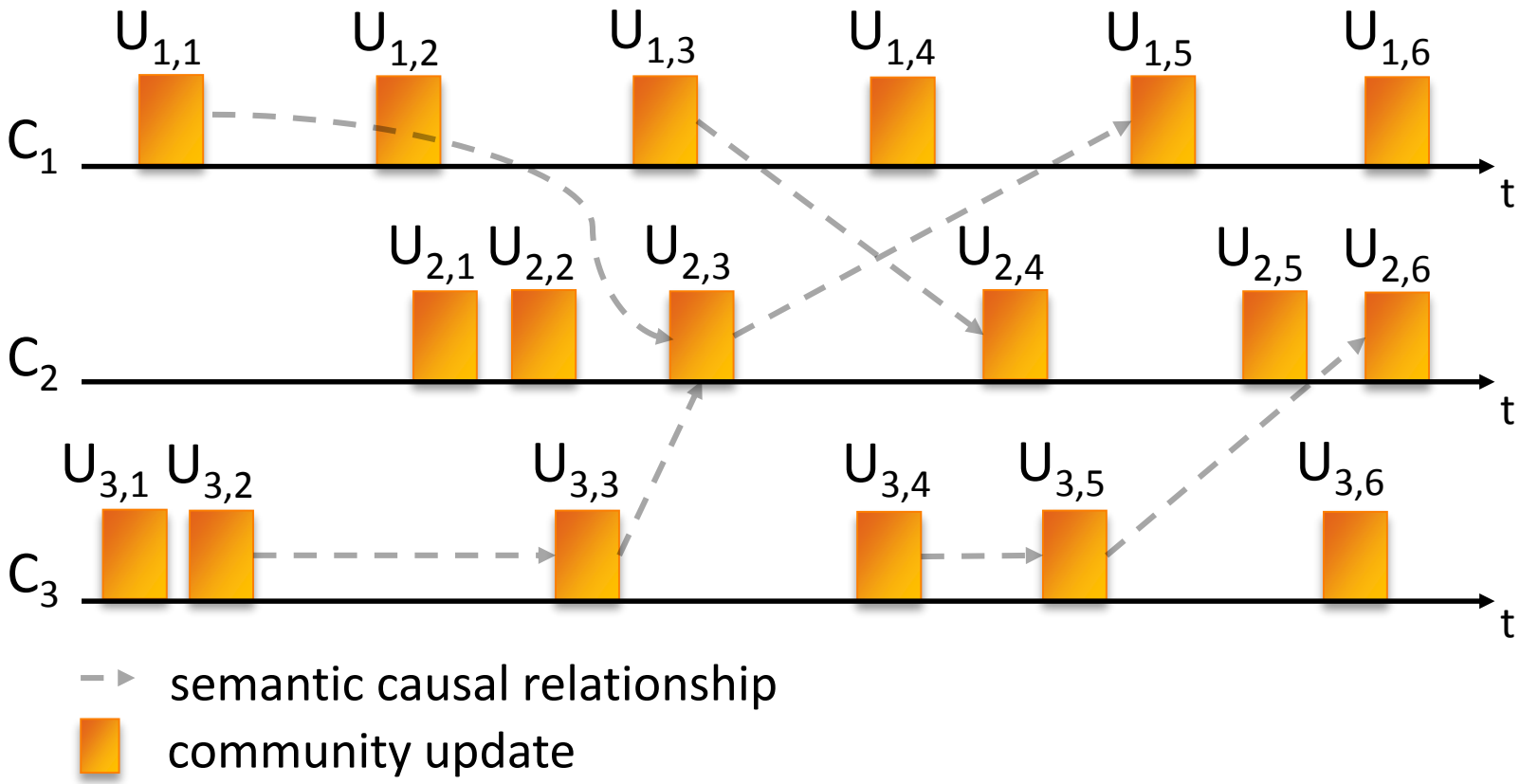
- Set of social communities $C = \{C_1 \dots C_n\}$
- each community C_i performs several updates $U_{i,j}$ with $j = 1 \dots m_i$ of the published information
- each published information can, in turn, influence updates of some other communities

Semantic Causal Relationships

Definition

It exists a *semantic causal relationship* between two updates $U_{x,y}$ and $U_{w,z}$ (with $x,w = 1 \dots n_i$ and $y,z = 1 \dots m_i$) **iff**:

- $U_{x,y}$ and $U_{w,z}$ are *semantically*-related
- $U_{x,y}$ and $U_{w,z}$ are *causally*-related


U1,1
U1,2
U1,3
U1,4
U1,5
U1,6
C1
t
U2,1
U2,2
U2,3
U2,4
U2,5
U2,6
C2
t
U3,1
U3,2
U3,3
U3,4
U3,5
U3,6
C3
t
semantic causal relationship
community update

Objectives of this work

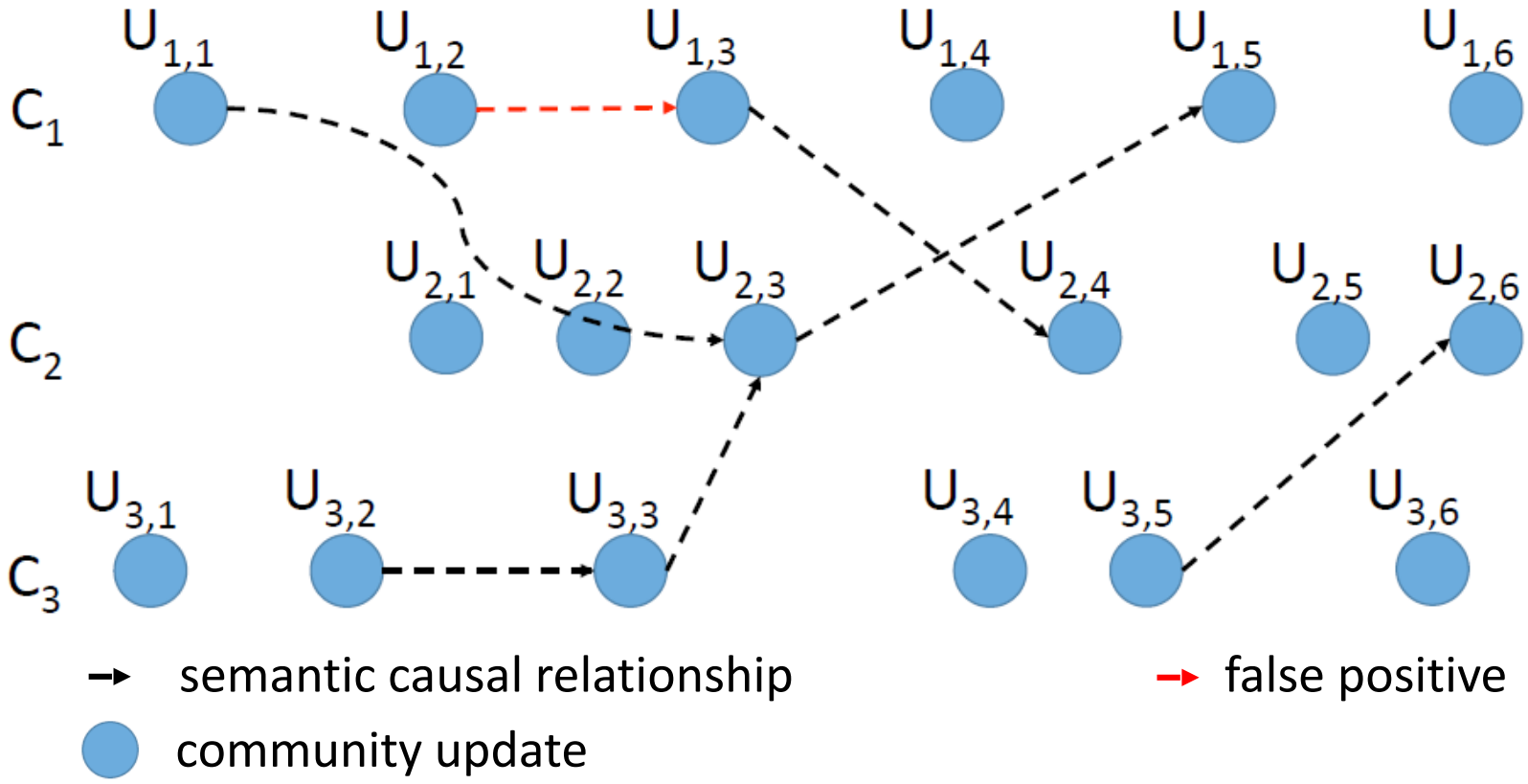
To propose an architecture to build a directed graph of semantic causal relationships

Formally, the built graph $\mathbf{G_s}$ is defined as:

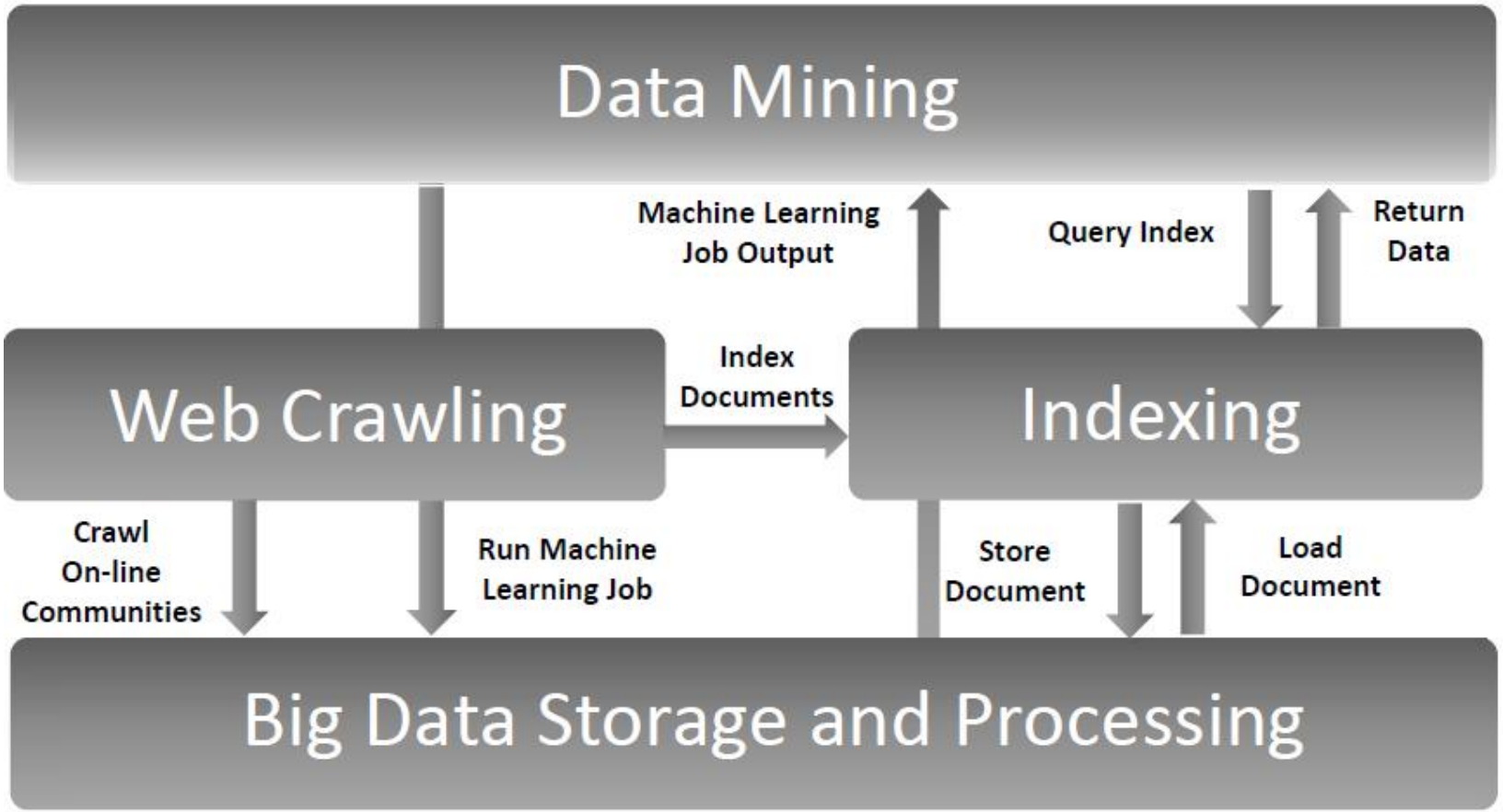
$$\mathbf{G_s : (V,E)\ ,\ E \subseteq \{\ (u,v)\ |\ u,v \in V}$$
$$\mathbf{u,v \text{ are } \textit{causally} \text{ and } \textit{semantically} \text{ related }\}}$$

V is the set of updates of social communities

E is the set of semantic causal relationships between updates


U1,1
U1,2
U1,3
U1,4
U1,5
U1,6
C1
U2,1
U2,2
U2,3
U2,4
U2,5
U2,6
C2
U3,1
U3,2
U3,3
U3,4
U3,5
U3,6
C3
semantic causal relationship
false positive
community update

Architecture Overview



Employed Technologies

- MapR
- Apache Solr
- Apache Nutch
- Apache Mahout

Why MapR ? (1/2)

- a production-ready distribution for Apache Hadoop developed by MapR Technolologies
- easy to use
- dependable
- especially fast
- Hadoop-API compatible

Why MapR ? (2/2)

Furthermore:

- includes a MapReduce module for parallel processing of large data sets
- supports the Hadoop distributed file system abstraction interface
- maintains compatibility with the Hadoop ecosystem and with the other Hadoop-related projects

Apache Solr (1/2)

- an open source enterprise distributed search platform
- highly reliable, scalable and fault tolerant
- able to support distributed indexing and distributed searching capabilities

All these capabilities are handled by a Solr sub-layer, called SolrCloud

Apache Solr (2/2)

The services offered by SolrCloud relies on *replication* and *sharding* techniques:

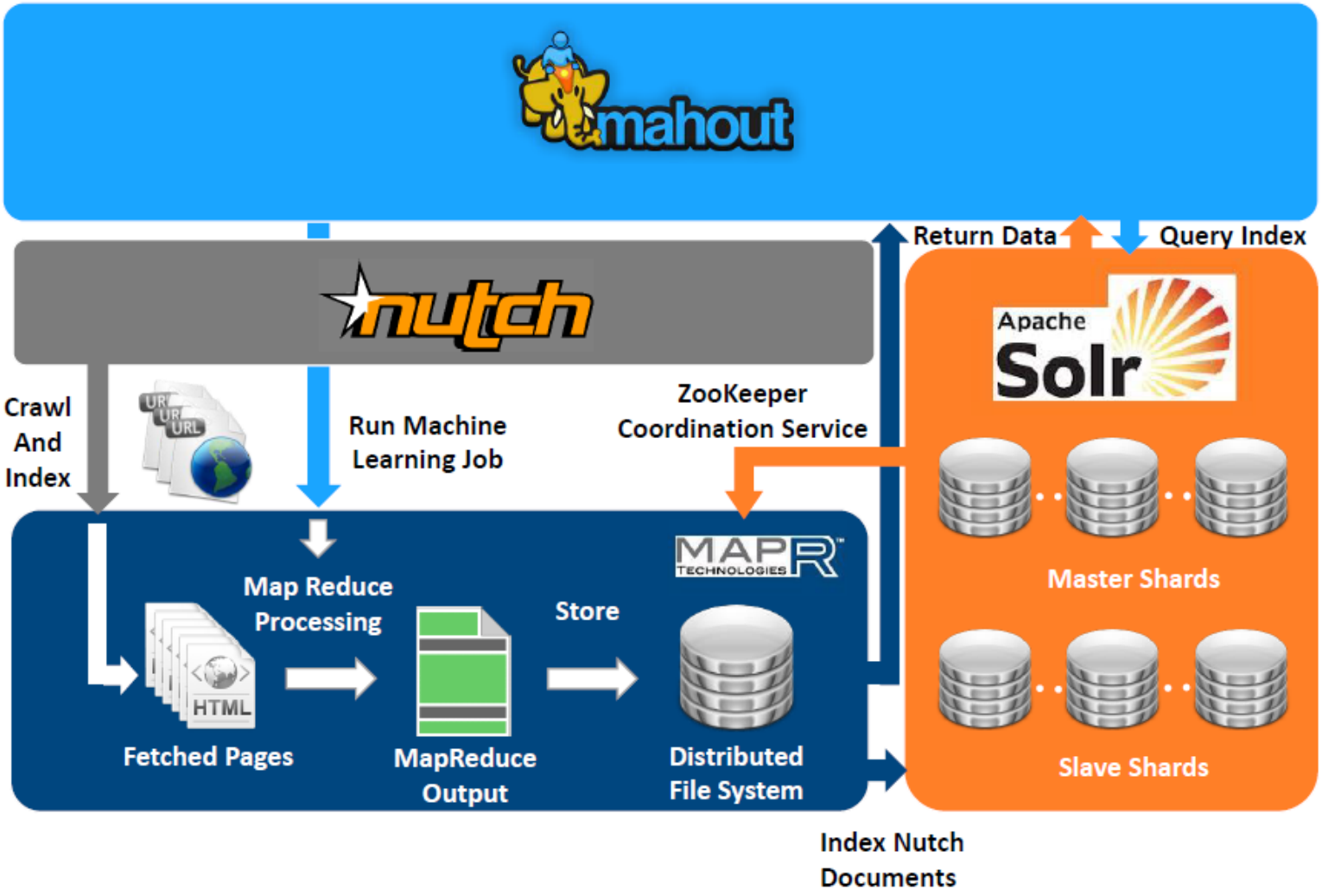
- *sharding* allows to split an index into multiple pieces, called *shards*
- *replication* ensures data redundancy and each index update can be issued to any shard

Apache Nutch & Apache Mahout

Apache Nutch is a highly extensible, robust and scalable open source crawler supporting the MapReduce paradigm. It observes politeness and implements a robot-exclusion protocol

Apache Mahout is a software library useful to produce free implementations of scalable machine learning and data mining algorithms

Technologies in the Architecture



Using Apache Mahout (1/2)

Once the updates of social communities have been successfully crawled:

- their content is both indexed in Solr and stored in the MapR-FS
- LSA (**L**atent **S**emantic **A**nalysis) is performed to establish semantic relationships between updates
- In Mahout, LSA is implemented through SSVD (**S**tochastic **S**ingle **V**alue **D**ecomposition) dimensionality reduction technique

Using Apache Mahout (2/2)

Hence, the architecture:

- exploits the Mahout framework to extract semantic relationships between updates using SSVD dimensionality reduction technique
- apply k-Means clustering to the extracted semantic relations
- post-process clustering results

Semantic Causal Relationship Graph Construction

After the clustering phase, the semantic graph is built through the following steps:

- within each cluster, the eventual duplicates of each social community's update are detected and deleted, maintaining only the oldest update's copy
- if after this preliminary phase, in a cluster, exist two or more updates referring to the same URL, the nodes representing such updates are included in the graph. These nodes are connected reflecting the causal relationship between them
- in an analogous way, similar pages belonging to the same cluster but to different domains are detected and added to the semantic graph

Detection of Similar and Duplicated Documents

Similar and duplicated documents detection is performed by using some similarity measure in information retrieval

Idea: to use the cosine similarity

Given two vectors and , respectively representing two documents and, the cosine similarity between these latter two is defined as:

$$similarity = \cos\vartheta = \frac{\overrightarrow{\mathbf{d}_1} \cdot \overrightarrow{\mathbf{d}_2}}{|\mathbf{d}_1|\,|\mathbf{d}_2|}$$

Cosine similarity is not enough!

INNOVATION INSIGHTS | community content | featured | blog

FOLLOW INNOVATION INSIGHTS

Should Your Startup's CEO Have Technical Chops?

BY TIM TUTTLE, EXPECT LABS 04.21.14 11:03 AM

Share 5 Tweet 18 +1 9 Share 9 Pin it

Are coding skills key to a startup's success? *Image: alexdecarvalho/Flickr*

MOST RECENT WIRED POSTS

Your Guide to Good IRL Behavior, From Vaping to Dressing Like a Techie

This Startup Says It Can Make Any Car Autonomous for $10,000

The Cold War Relics Three Photographers Are Documenting Before They Disappear

Amazon's Fire Phone May Be Too Magical for Its Own Good

Cosine similarity is not enough!

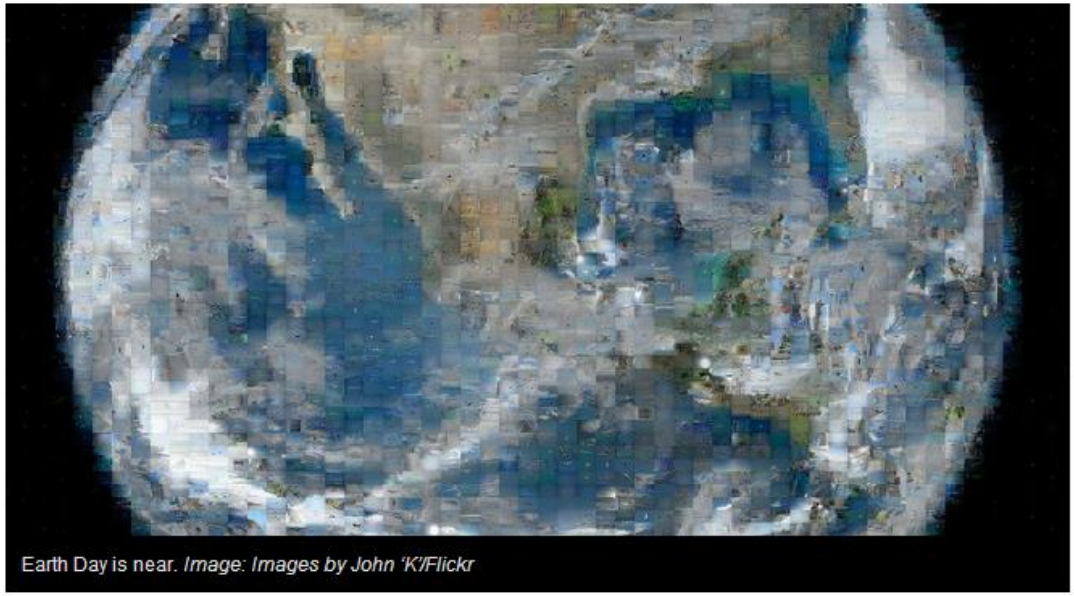
INNOVATION INSIGHTS | community content | featured | blog

FOLLOW INNOVATION INSIGHTS

Revving Up IT Performance by Recycling

BY DEREK BRITTON, MICRO FOCUS 04.21.14 11:03 AM

Share 2 Tweet 28 +1 4 Share 25 Pin it



Earth Day is near. *Image: Images by John 'K'/Flickr*

MOST RECENT WIRED POSTS

A Sleek New Hearing Aid That Solves a Nagging Problem

Volcano World Cup: Group G

Does the Introductory Physics Course Cover Too Much?

"Individual Actions Are Doomed to Failure": Coalition Asks for Global Action on Antibiotics

Case Study (1/2)

Three technology news and information websites:

- Ars Technica
- Engadget
- Wired

These three communities have been crawled continuously for a period of about two weeks

Case Study (2/2)

The crawling is performed by a five-node MapR cluster, composed of:

- 3 Nutch crawlers
- 2 Solr shards (no replication)

The majority of the nodes runs Zookeeper coordination services. The Mahout machine learning jobs are executed on three nodes

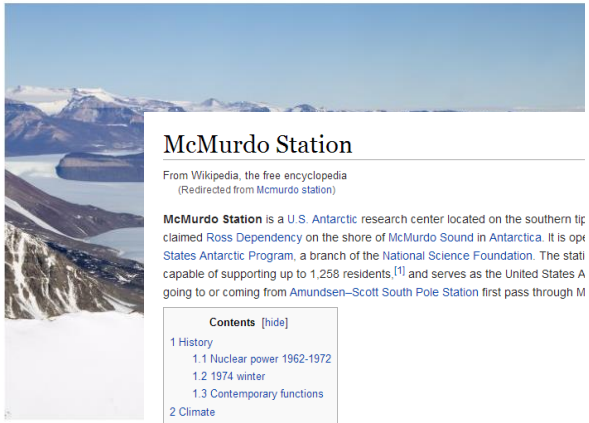
Mahout Clustering Output Sample

Using radioactive krypton to find ancient glacial ice

Not kryptonite in the Fortress of Solitude—the noble gas trapped in air bubbles.

by **Scott K. Johnson** - Apr 21 2014, 9:00pm E

EARTH SCIEN



Taylor Glacier in Antarctica

NASA/Earth Observatory/Ro

There's a reason that
Two-Mile Time Machine
of Earth's climate and
easier when you can p
an inconveniently large
Machine" is that it's on
years, and the oldest (

If we could go a little fu
terribly interesting clim
after, we've experience
cores, but they can't pr

McMurdo Station

From Wikipedia, the free encyclopedia
(Redirected from Mcmurdo station)

McMurdo Station is a U.S. Antarctic research center located on the southern tip
claimed Ross Dependency on the shore of McMurdo Sound in Antarctica. It is ope
States Antarctic Program, a branch of the National Science Foundation. The stati
capable of supporting up to 1,258 residents,[1] and serves as the United States A
going to or coming from Amundsen–Scott South Pole Station first pass through M

Contents [hide]

1 History
- 1.1 Nuclear power 1962-1972
- 1.2 1974 winter
- 1.3 Contemporary functions

2 Climate
3 Communications
4 Transportation
- 4.1 Air
- 4.2 Surface

5 Historic sites
6 Points of interest
7 In popular culture
- 7.1 Literature
- 7.2 Film and television

8 See also
9 Notes
10 References
11 External links

History [edit]

The station owes its designation to nearby McMurdo Sound, named after Lieutenant Archibald McMurdo of H.M.S. *Terror*, which first charted the area in 1841 under the command of British explorer James Clark Ross. British explorer Robert Falcon Scott first established a base close to this spot in 1902 and built Discovery Hut, still standing adjacent to the harbour at Hut Point. The volcanic rock of the site is the southernmost bare ground accessible by ship in the Antarctic. The United States officially opened its first station at McMurdo on Feb. 16, 1956. Founders initially called the station *Naval Air Facility McMurdo*.

SCOTT K. JOHNSON / ASSOCIATE WRITER

er shrank quickly in the past

oned a repeat performance.

EMAIL

RECENT STORIES BY SCOTT K. JOHNSON

Taylor Glacier

From Wikipedia, the free encyclopedia

Coordinates: 77°44'S 162°10'E

This article is about the glacier in Antarctica. For the glacier in Colorado, USA, see Taylor Glacier (Colorado).

The **Taylor Glacier** is an Antarctic glacier about 54 kilometres (34 mi) long, flowing from the plateau of Victoria Land into the western end of Taylor Valley, north of the Kukri Hills, south of the Asgard Range. The middle part of the glacier is bounded on the north by the Inland Forts and on the south by Beacon Valley.

The glacier was discovered by the British National Antarctic Expedition (1901–04) and at that time thought to be a part of Ferrar Glacier. The Western Journey Party of the British Antarctic Expedition 1910 determined that the upper and lower portions of what was then known as Ferrar Glacier are apposed, i.e., joined in Siamese-twin fashion north of Knobhead. With this discovery Scott named the upper portion for Griffith Taylor, geologist and leader of the Western Journey Party.

The Taylor Glacier has been the focus of a measurement and modeling effort carried out by researchers from the University of California at Berkeley and the University of Texas at Austin.

Like other glaciers in the McMurdo Dry Valleys, Taylor Glacier is "cold-based," meaning its bottom is frozen to the ground below. The rest of the world's glaciers are "wet-based," meaning they scrape over the bedrock, picking up and leaving obvious piles c

Cold-based glaciers flow m
debris, cause little erosion,
surfaces full of crevasses,

Taylor Glacier

Type Polar Glacier
Location Antarctica

References [edit]

1. ^ Taylor Valley, Antarc
this US government we

This article incorporates
document "Taylor Glacier"

See also [edit]

- Blood Falls, an escapin
Glacier over a million ye
- List of glaciers
- List of glaciers in the Ar

McMurdo Station from above

74
ENGADGET SCORE

ASUS
Zenbook Prime UX31A

19 more versions

AS LOW AS
$1089

GET PRICE ALERTS

REVIEWS • 25 SPECS ANSWERS • 7 DISCUSSIONS • 5

share:

add to compare enlarge

A high-res display, and a much-improved keyboard

74

REVIEW BY SARAH SILBERT

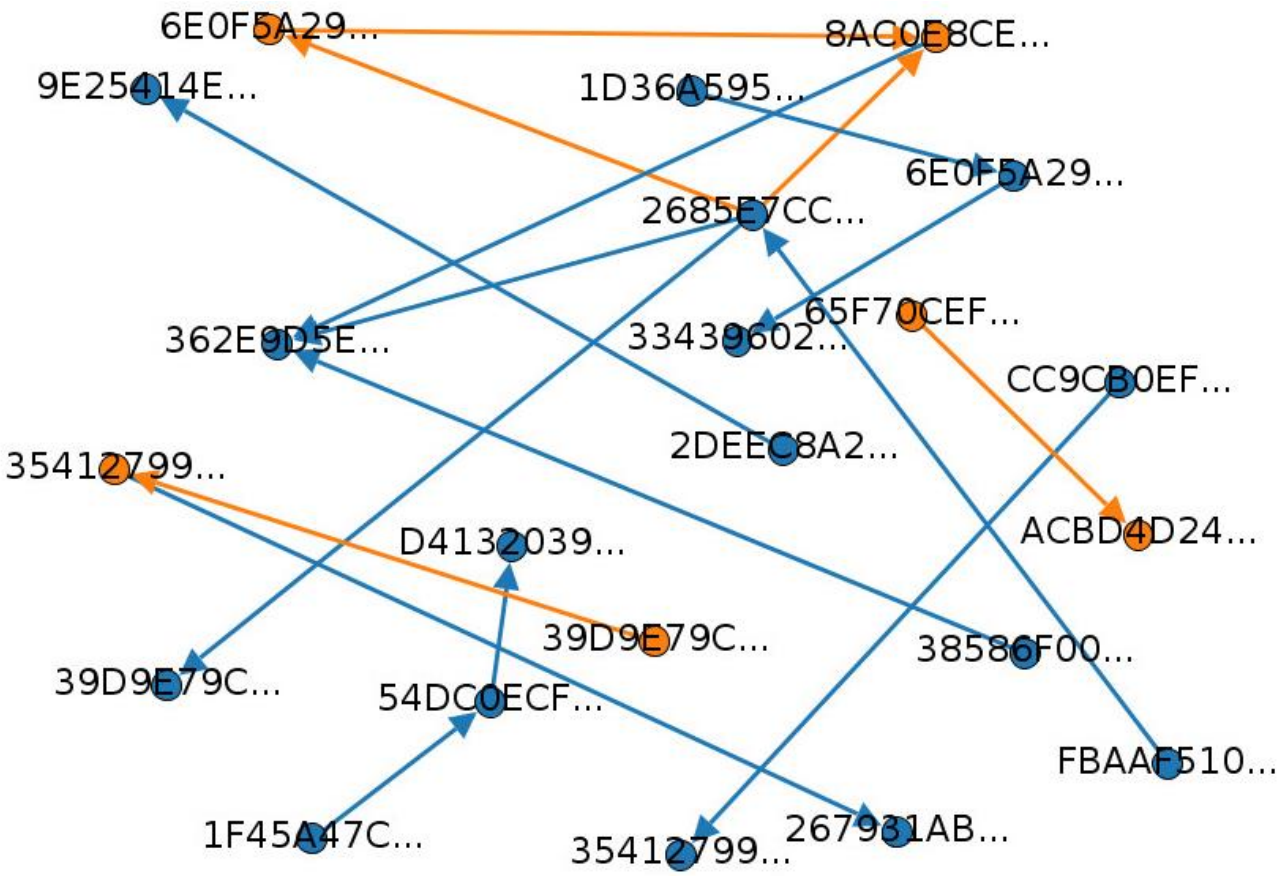
a week ago

When ASUS first released the UX31E last fall, we found plenty to like in its striking design, high-quality display and brisk performance. The addition of backlighting and a more comfortable keyboard -- not to mention the step up to Ivy Bridge and Intel HD Graphics 4000 -- makes the whole package considerably better.

But that doesn't mean the Zenbook Prime UX31A is now the best. To claim that title, it needs a more usable trackpad, above all else. We still think the MacBook Air is a better all-around ultraportable, as it offers a more comfortable touchpad and keyboard in a similarly attractive package, though we wish it packed an IPS 1080p display like the one on the UX31A. And though it's considerably more expensive, you might also be happier with the Samsung Series 9, which lasts longer on a charge, rocks an impressive display of its own and sports a more reliable trackpad. Still, given all the UX31A has to offer, chances are you'll be pleased with your purchase. Just evaluate your patience for temperamental touchpads before you pull the trigger.

Read our full review →

Semantic Causal Relationship Graph Sample



Conclusions

- on-line social communities are a fundamental source of information in business and information security intelligence
- an architecture to extract semantic causal relationships between updates of social communities has been presented
- most of the architecture relies on open source frameworks
- detection of similar and duplicated documents needs a pre-processing phase
- additional tools can use the semantic causal relationship graph, as input, for further investigations

Questions ?

CIS Sapienza

Research Center for Cyber Intelligence and information Security