

The Observatory – The structure of
news: monitoring online media topics
with mutual information

David M.S. Rodrigues

Track D - Complexity and Computer Science

ECCS'10 - Lisbon, 13-17 September 2010

<http://theobservatorium.eu/>

Summary



- The Problem
- Our Approach at *The Observatorium*
- Results and Conclusions

I. The Problem

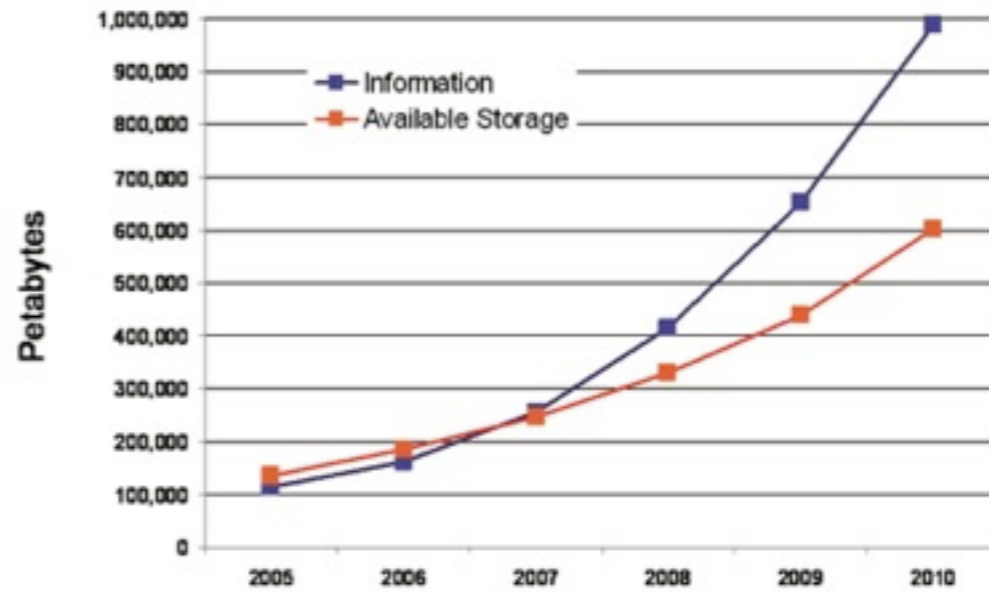


"The universe (which others call the Library) is composed of an indefinite and perhaps infinite number of hexagonal galleries, with vast air shafts between, surrounded by very low railings. From any of the hexagons one can see, interminably, the upper and lower floors."

-- The Library of Babel, Jorge Luis Borges

Figure 2

Information Versus Available Storage




Source: IDC, 2007

[IDC Report “The Expanding Digital Universe”, 2007]





 Everything

 Images

 Videos

 More

All results

Social

 More search tools

ECCS'10

eccs'10

eccs'

About 264,000 results (0.24 seconds)

[Welcome to the main conference website | ECCS'10 European ...](#) ☆

ECCS'10 will be located at the Lisbon University Institute, from the 13th to 17th September, 2010. The main conference tracks will be presented on the 13th, ...

www.eccs2010.eu/ - [Cached](#)

[ECCS'10 Challenge | ECCS'10 European Conference on Complex Systems](#) ☆

The results of the Challenge will be published in the **ECCS'10** webpage and in the ASSYST Newsletter. Also, a detailed report concerning the **ECCS'10** Challenge ...

www.eccs2010.eu/challenge - [Cached](#)

 [Show more results from www.eccs2010.eu](#)

The Observatorium

Real-time monitoring of multi-level network structures for the study of knowledge generation and opinion dynamics in the Internet

Internet has become the main medium for social systems that interactively exchange ideas and opinions, generating and sharing knowledge worldwide. Networks such as communication and social networks are used to inform and exchange arguments, generating opinions concerning socially relevant domains, such as politics, economics or culture. These knowledge generation systems are sustained by blogs, news web pages, opinion articles proposed by journalists, politicians, and other interactive network structures. Such systems are characterized by its constant change in the way they appear, evolve, and disappear being substituted by new communication structures. Knowledge and opinion systems are complex given the intricacy of the different levels of network structures participating in its dynamics, such as communication networks, social networks and knowledge networks composed by topics and linguistic concepts. These networks are all composed of many interconnected and interdependent components, which grow without centralized control concerning the physics of information diffusion.



THE OBSERVATORIUM

[Start](#)

[Theseus](#)

[Team](#)

[Working Papers](#)

[Software](#)

[ECCS 2010 Conference](#)

[Complex Systems Studies](#)

- [Show pagesource](#)
- [Old revisions](#)
- [Recent changes](#)
- [Backlinks](#)
- [Index](#)
- [Login](#)

theobservatorium.eu

I. The Problem



- Challenge: understand how and what are people discussing, what are their interests, subjects, opinions, arguments, communication structure
- Difficulty: Elusiveness of information.
- Goal: characterise opinion dynamics extracted and deduced from large and diverse data

EL PAÍS.com website screenshot. The main headline is "El Gobierno español rebaja un 5% el sueldo de los funcionarios". Other visible headlines include "El Supremo resuelve la causa contra Camps por los trajes que le regaló la trama Gürtel" and "Bruselas afirma que 'van en la buena dirección'". The page features a navigation menu, a search bar, and several article teasers with images.

Le Monde.fr website screenshot. The main headline is "La TV 3D arrive enfin dans votre salon". Other visible headlines include "L'Europe a repris la main" and "Comment reconnaître le vrai 'agitateur' du fax". The page features a navigation menu, a search bar, and several article teasers with images.

guardian.co.uk website screenshot. The main headline is "Labour set to pounce if Lib Dem-Toy talks fail". Other visible headlines include "LIVE: EU financial crisis" and "KBS sets 2,600 more jobs". The page features a navigation menu, a search bar, and several article teasers with images.

The Australian website screenshot. The main headline is "Dragan captured after 43 days on run". Other visible headlines include "BUDGET 2010" and "Tu és o visitante 999.999!". The page features a navigation menu, a search bar, and several article teasers with images.

Politika website screenshot. The main headline is "Шок и ужас в Грција". Other visible headlines include "Вреќа и сликата на граѓанската журналистика" and "Божките знамена на Бугария". The page features a navigation menu, a search bar, and several article teasers with images.

P20 website screenshot. The main headline is "Ber compelo vale directamente 13 a 20 milhões ao Benfica". Other visible headlines include "Governo abre a porta a aumentos de impostos" and "Papa em Portugal". The page features a navigation menu, a search bar, and several article teasers with images.

2. Our Approach



- Automatic categorisation of text from linguistics, natural language, statistics and information retrieval.
Strategies:
 - regression models
 - Bayesian approaches
 - nearest neighbour classification
 - neural networks
 - hierarchical clustering

[Miao and Qiu, 2010] [Solé et al., 2010]



2. Our Approach



- Automatic categorisation of text from linguistics, natural language, statistics and information retrieval.

Strategies:

- regression models

- **Human intervention!**
Bayesian approaches

- nearest neighbour classification

- neural networks

- hierarchical clustering

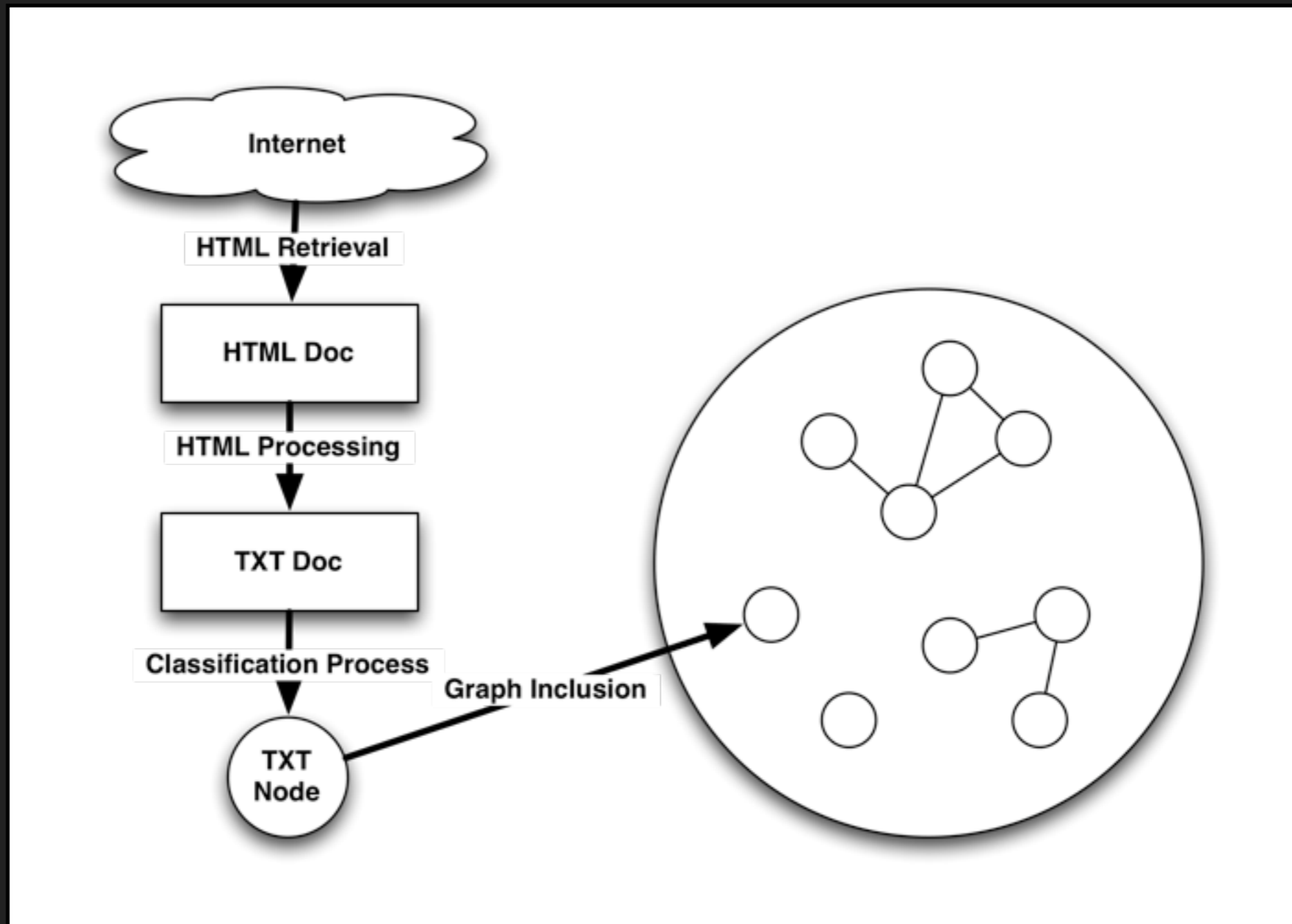
[Miao and Qiu, 2010] [Solé et al., 2010]



Monitoring topic trends from on-line media



General methodology for acquiring and processing text documents from the Internet





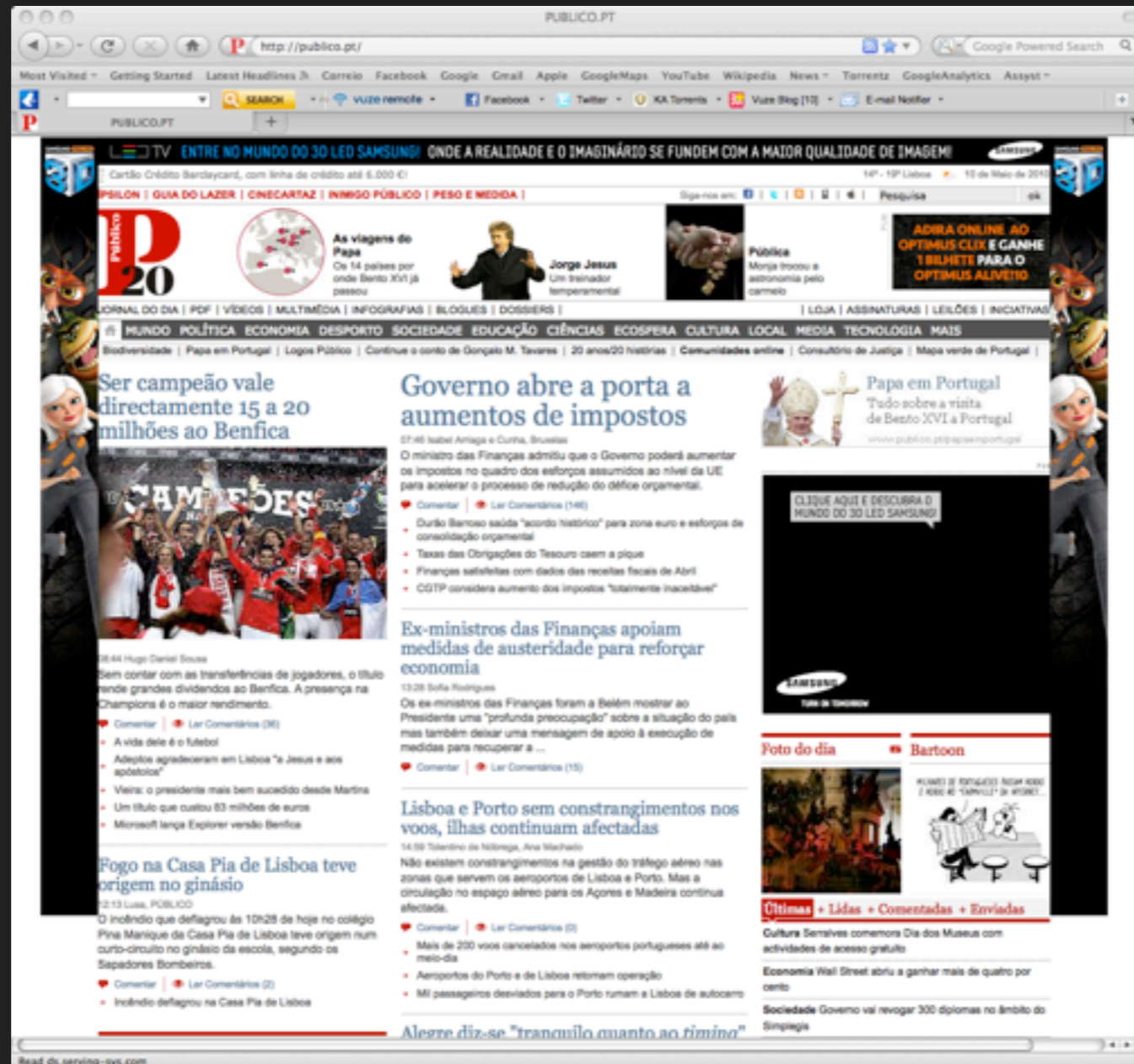
1st experiments - using on-line newspapers:

1. Retrieval of on-line newspapers
2. Analysis of news items (classification)
3. Representation of topic networks
4. Real-time monitoring of the dynamics of topic networks

Monitoring topic trends from on-line media



Extraction of the text from each HTML file by employing the Text to Tag ratio proposed in (Weninger, 2008)



<http://www.politika.bg/article?id=17018>

Monitoring topic trends from on-line media



Extraction of the text from each HTML file by employing the Text to Tag ratio proposed in (Weninger, 2008)

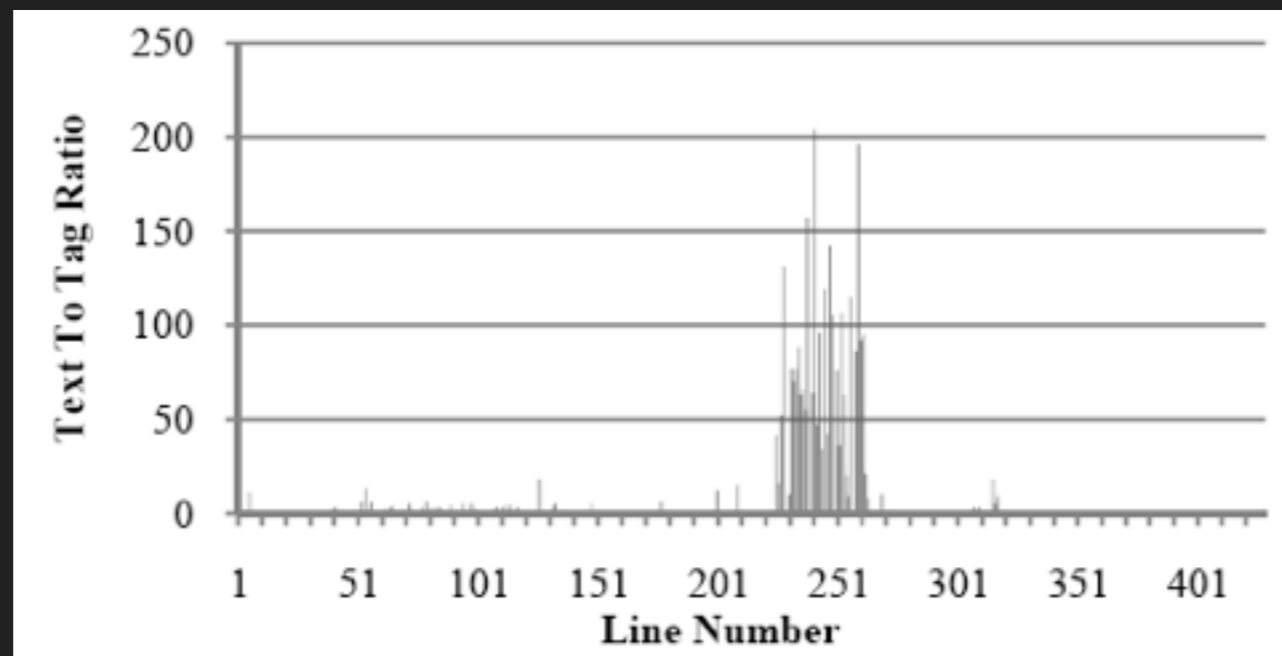
```
article.html
<div class="menuContainer">
<div class="mnavigationBox">
  <div class="mnavigation"><ul>
    <li><a href="/">главна</a></li><li><a href="/theme">Темата</a></li><li><a href="/authority" class="activebefore">
  >Властта</a></li><li><a href="/opinions" class="active">Мненията</a></li><li><a href="/world">Светът</a></li><li><a
  href="/reference">Игрите</a></li><li><a href="/bussiness">Бизнесът</a></li><li><a href="/society">Животът</a></li><li><a
  href="/interview">Интервю</a></li><li><a href="/articles?id=102">Образование</a></li><li><a
  href="/articles?id=103">Скорост</a></li> </ul>
  </div>
</div>
</div>
</div>
<script type="text/javascript" src="/js/jquery.js"></script>
<script>
function changeMat(val){
  document.location="materialView?id=" + val;
}
</script>
<div class="container">
<table cellspacing="0" cellpadding="0">
<tr><td width="640" valign="top">
<div id="articleContainer">
  <div class="articleTitle">Олер Йорданов: Искам го този Ейфеловокул</div>
  <div class="articleSubtitle"></div>
  <div style="float:left; margin: 10px 0px 10px 5px; overflow:hidden;clear:right;display:inline">
  <br>
  </div>
  <div class="articleAuthorDateBox">
  <div style="width: 620px;float:left; padding-bottom: 20px;"><span class="articleDate">30.04.2010 </span></div>
  Според психолози страхът от катастрофи е специфичен симптом на хронична тревожност, при която хората, страдащи от това
  - разстройство, са спомождени постоянно от ирационални мисли. Търся това определение, след като на опашката дочувам: „Не можеш да
  - се запасиш достатъчно, вулканът ще отрови всички“. Познати от туристическия бизнес се обаждат разтревожени да ме питат ще летят
  - ли изобщо самолети това лято. Пак според психолозите хората определено се страхуват от неизвестното повече, отколкото от
  - познатото. При станалата вече норма ниска природно-научна култура на българите, вълните катастрофични психози изглеждат почти
  - предопределени. Ето защо сядам да напиша нещо за вулканите. Не като опит за масова терапия. Четящите и знаещи хора едва ли се
  - нуждаят от нея.<br><br>В ежедневната ни представа Земята е нещо твърдо и солидно. Но като изключим една кора с дебелина от 8 км
  - (при океанските дъна) до около 70 км (при континентите), Земята в недрата си изобщо не е твърда. Първите 2900 км в дълбочина
  - представлява гореща пластинна субстанция, която геофизиците наричат мантия (обвивка). В центъра е желязно-никелово ядро с радиус
  - цели 3478 км, но и то в по-голямата си външната си част е течно.<br><br>Как знаем всичко това след като никой не е бил там,
  - нито пък е възможно пряко наблюдение? За да узнаят структурата на земята, учените се възползват от едни от най-разрушителните
  - природни явления. Всяко земетресение става източник на вълни, подобни, но и различни на вълните, които се предизвикват от хвърлен
  - във вода камък. Тези вълни се наричат сеизмични и се разпространяват през цялото земно кълбо. Веднъж регистрирани от множество
  - станции, включително от такива на обратната страна на глобуса, въпрос е само на (не проста) математика да се определи видът,
  - структурата и температурата на средите, през които са преминали.<br><br>Като кажем „крехко“ си представяме яйце. Но ако кората на
  - Земята беше толкова тънка, колкото тази на яйцето, съотносително на размерите им, тя трябва да е поне пет пъти по-дебела. Т. е.
```



Monitoring topic trends from on-line media



Extraction of the text from each HTML file by employing the Text to Tag ratio proposed in (Weninger, 2008)



(Weninger, 2008)



Monitoring topic trends from on-line media



Extraction of the text from each HTML file by employing the Text to Tag ratio proposed in (Weninger, 2008)





Analysis of news: topic networks

- Each news item is a node of the Graph
- A node is linked to other nodes according to a the Jaccard distance on words.
- The Jaccard similarity coefficient is used to calculate distances

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



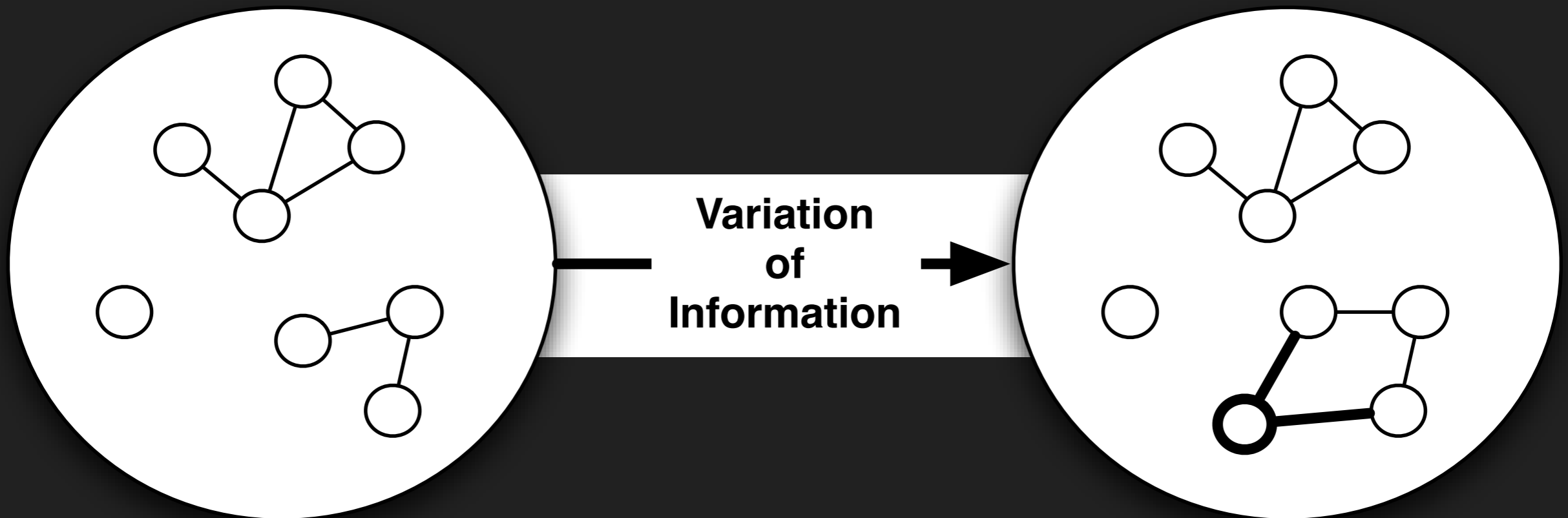
Monitoring of the dynamics of topic networks

The life expectancy of a node to graph G is given by the *time to live* (TTL) of this node

Each time we added a node to graph G , the TTL of the nodes that have established connections to it is incremented, and for others this value is decremented



Variation of information (Meilã, 2007)



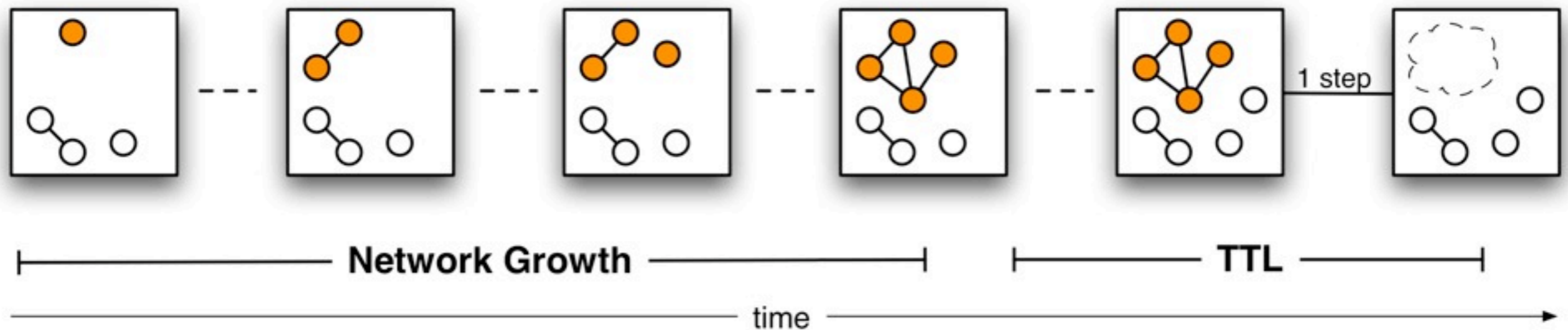
$$I(C, C') = \sum_{k=1}^K \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P(k')}$$

$$VI(C, C') = H(C) + H(C') - 2I(C, C')$$



Topic detection

Schematics of topic growth and detection via VI



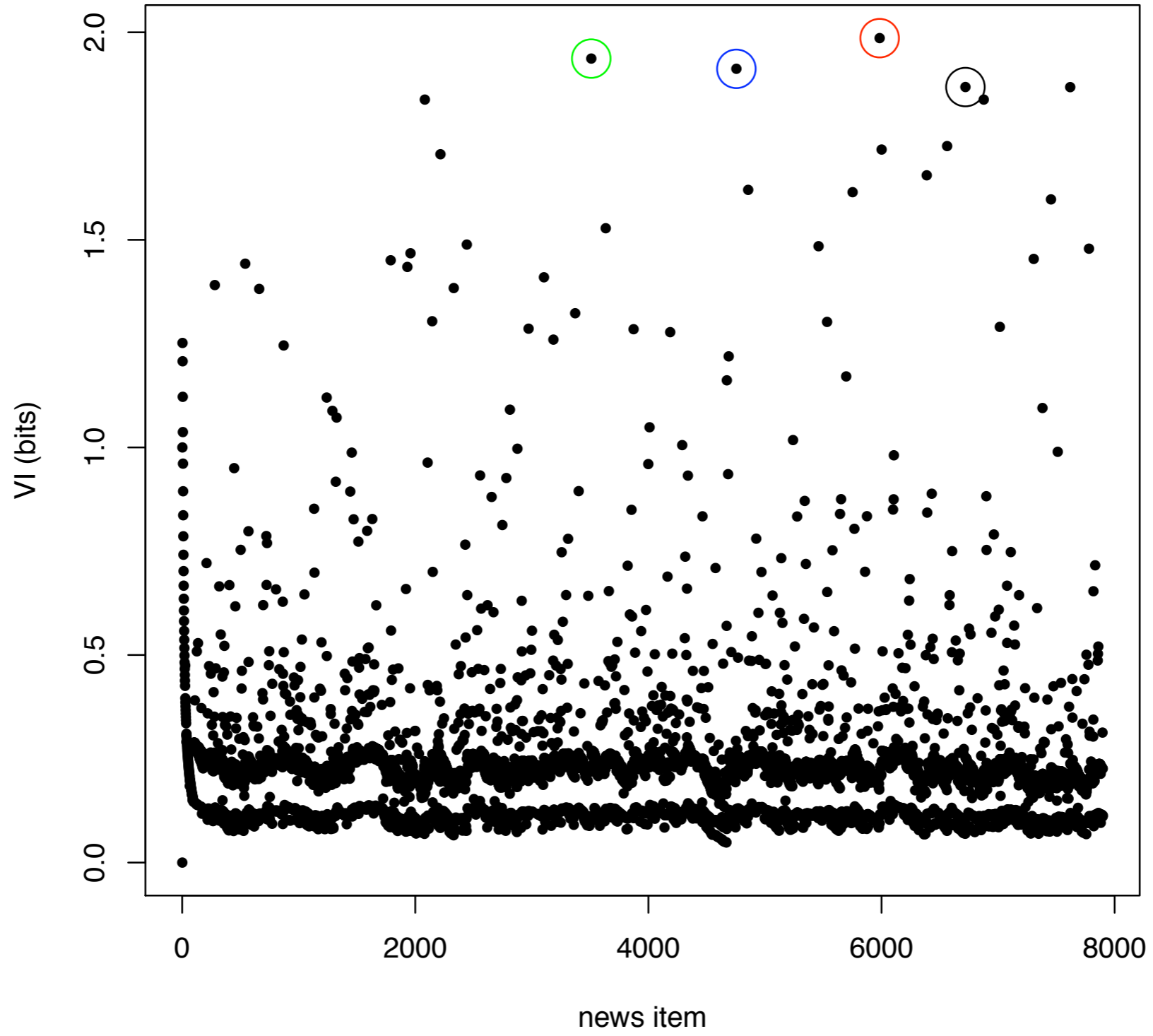
3. Results

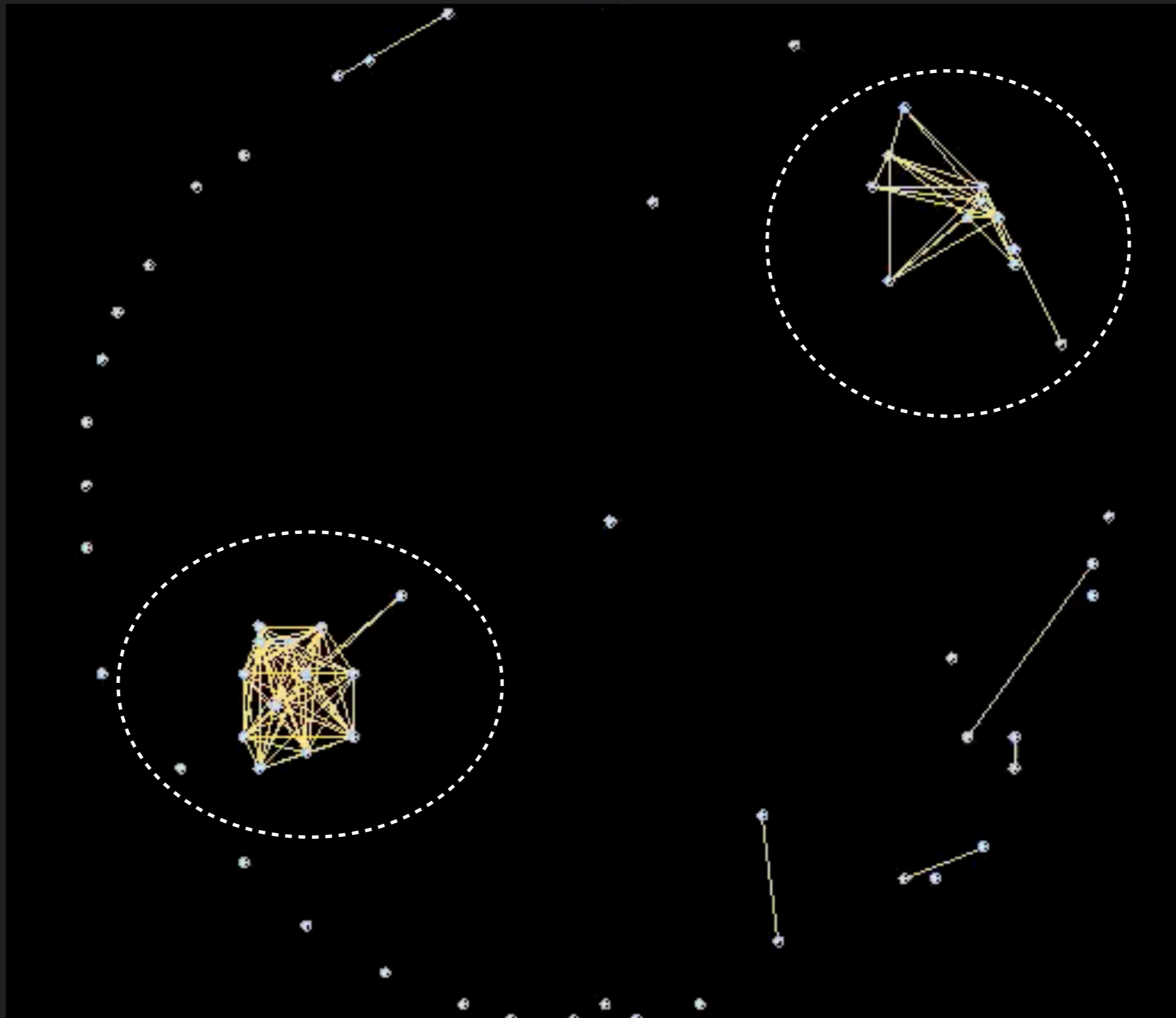


Example: 7928 news items collected from the *Público* newspaper from Nov. 11, 2009 to Jan. 25, 2010



Variation of Information





3510

“payment” “international” “Portugal” “stock market” “study”
“negotiation” “comments”

4755

“next” “comments” “industrial” “one hundred” “trust” “export”

5984

“Madrid” “tournament” “Sporting” “football player” “players”
“game” “comments”

6720

“international” “presentation” “comments” “prototype” “press”
“development” “american”

3. Conclusions



- Versatile method
- Simultaneous Extraction and Classification
- No *a priori* knowledge*



Next:

- general rules to eliminate irrelevant words
- deploy the crawlers to social sites like blogs, twitter, facebook, myspace, etc... and not just newspapers
- do multi-channel correlation and topic extraction, meaning that it will not be done only across time, but also across “web space” in different domains

Monitoring topic trends from on-line media



- on-line software for free usage, with datasets
- deploy the online tools for researchers to use them in an easy friendly way. Also, software will be available under open source licence

- any questions?

Monitoring topic trends from on-line media



- on-line software for free usage, with datasets
- deploy the online tools for researchers to use them in an easy friendly way. Also, software will be available under open source licence

<http://theobservatorium.eu/>

- any questions?