The Observatorium — 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

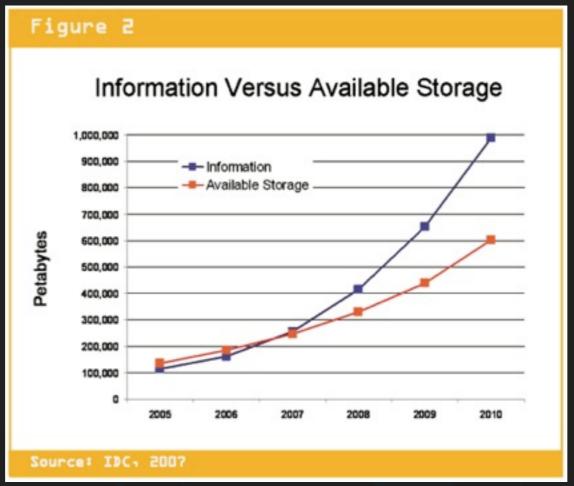


- The Problem
- Our Approach at The Observatorium
- Results and Conclusions

1. The Problem

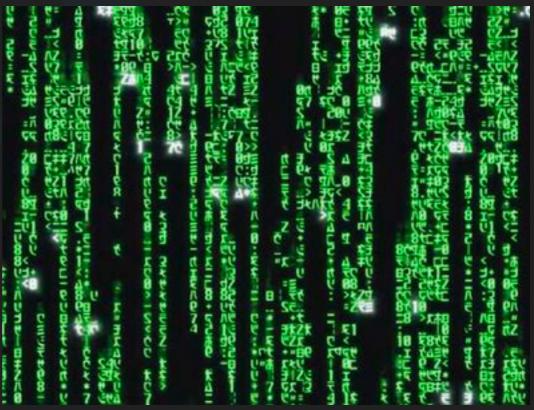
"The universe (which others call the Library) is composed of an indefinite and perhaps infinite number of hexagonal galleries, with vast air shafts between, surrounded by very low railings. From any of the hexagons one can see, interminably, the upper and lower floors."

-- The Library of Babel, Jorge Luis Borges

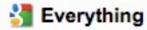


[IDC Report "The Expanding Digital Universe", 2007]













▼ 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

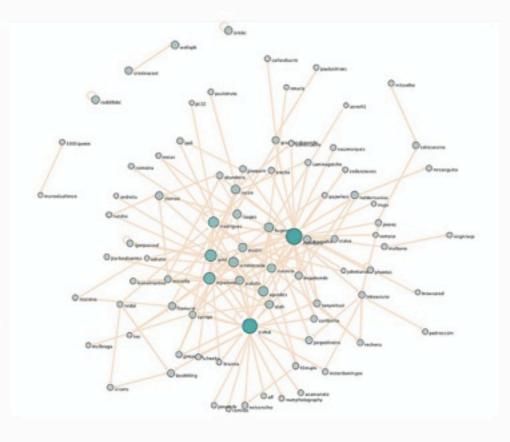
The results of the Challenge will be published in the ECCS'10 webpage and in the ASSYST Newsletter. Also, a detailed report concerning the ECCS'10 Challenge ... www.eccs2010.eu/challenge - Cached

◆ 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

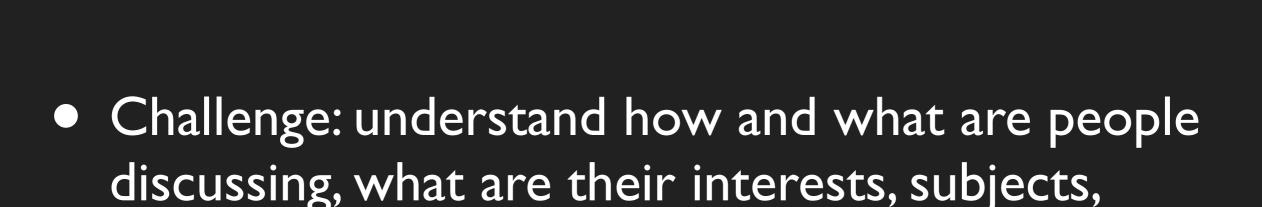
ECCS 2010 Conference

Complex Systems Studies

- Show pagesource
- Old revisions
- = Recent changes
- Backlinks
- Index
- Login

theobservatorium.eu

I. The Problem



opinions, arguments, communication structure

- Difficulty: Elusiveness of information.
- Goal: characterise opinion dynamics extracted and deduced from large and diverse data













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

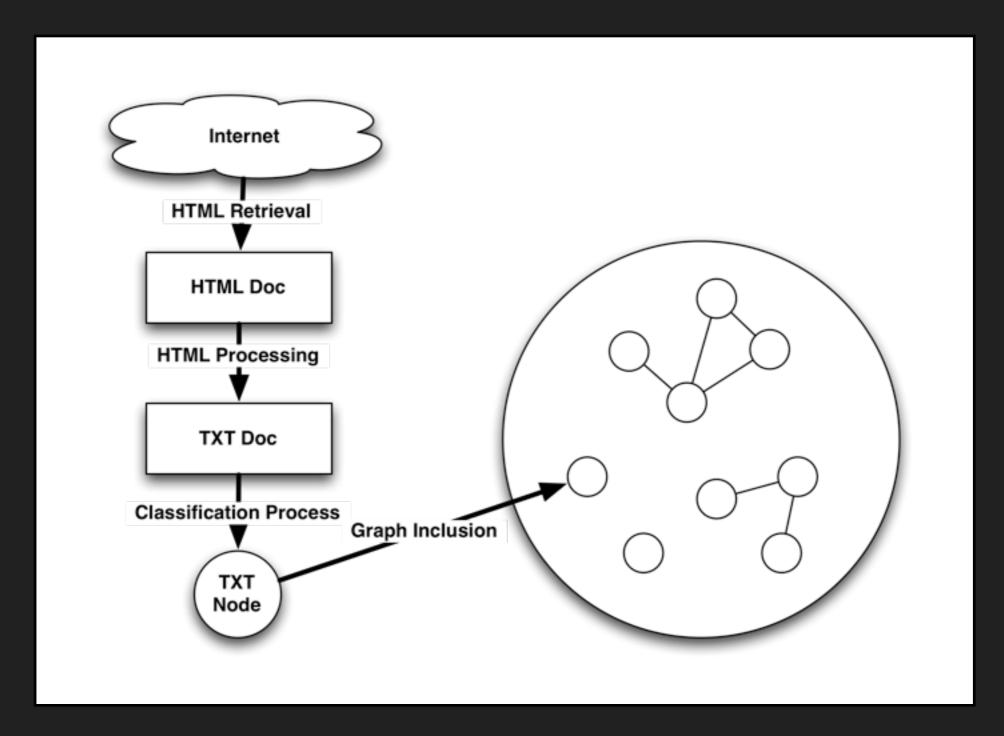


- Automatic categorisation of text from linguistics, natural language, statistics and information retrieval.
 Strategies:
 - regression models
 - Bayesian approaches
 - nearest neighbour classification
 - neural networks
 - hierarchical clustering

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



General methodology for acquiring and processing text documents from the Internet





Ist experiments - using on-line newspapers:

- I. 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



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

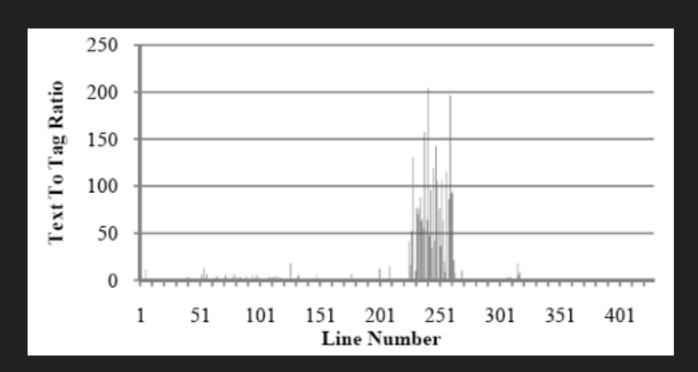


Extraction of the text from each HTML file by employing the Text to Tag ratio proposed in (Weninger, 2008)

```
odiv class="menuContainer";
   <div class="mnavigationBox">
         <div class="mnavigation">
    «li»«a href="/">Γπαβκα</α>«li»«a href="/theme">Τεκατα</α>«a href="/authority" class="activebefore"
   >Властта</a>«a href="/opinions" class="active" >Мненията</a>«a href="/world">Светьт</a>«li>«а
   href="/reference">Mrpaчитe</a>\absolute{\absolute} \ref="/bussiness">5изнесьт</a>\absolute{\absolute} \ref="/society">Животът</a>\absolute{\absolute} \ref="/society">Животът</a>\absolute
   href="/interview">Mxrepso</a><a href="/articles?id=182">Obpasosanxe</a><a href="/articles?id=182">Obpasosanxe</a><a href="/articles?id=182">Obpasosanxe</a></a>
  href="/articles?id=103">Cxopoct</a>
FI </div>
□ </div>
O </div>
   <script type="text/javascript" src="/js/jquery.js"></script>
function changeMat(val){
         document.location="materialView?id=" + val;
</script>
adiv class="container">
articleContainer"
               <div class="articleTitle">Олег Йорданов: Искам го този Ейяфялайокул!</div>
         <div class="articleSubtitle"></div>
                         <div style="float:left; margin: 10px 0px 10px 5px; overflow:hidden;clear:right;display:inline">
                    <img src="/img/?id=31360_sz=0_cut=yes" id="artImg"><br>
         <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 спохождани постоянно от ирационални мисли. Търся това определение, след като на опашката дочувам: "Не можеш да
   се запасиш достатъчно, вулканът ще отрови всичко". 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 сядом да напиша нещо за вулканите. Не като опит за масова терапия. Четящите и знаещи хора едва ли се
   нуждаят от нея.∢р-</р⊳В ежедневната ни представа Земята е нещо твърдо и солидно. Но като изключим една кора с дебелина от 8 км
   (при океанските дъна) до окало 70 км (при континентите), Земята в недрата си изобщо не е твърда. Първите 2900 км в дълбочина
   представлява гореща пластична субстанция, която геофизиците наричат мантия (обвивка). В центъра е желязно-никелово ядро с родиус
   щели 3470 км, но и то в по-голямата си външната си част е течно.
   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 жвърлен
   във вода комък. 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. «р. «/р. Като кажем "крехко" си представяме яйце. Но ако кората на
    Земята беше толкова тынка, колкото тази на яйцето, съотносително на размерите им, тя трябва да е поне пет пъти по-дебела. Т. е.
                                                      ↑ O v Tab Size: 4 ↑ -
       1 Column: 1 3 HTML
```



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



(Weninger, 2008)



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

Според психолози страхът от катастрофи е специфичен симптом на хронична тревожност, при която хората, страдаци от това разстройство, са спохождани постоянно от ирационални мисли. Търся това определение, след като на опашката дочувам: "Не можеш да се запасиш достатьчно, вулканът ще отрови всичко". Познати от туристическия бизнес се обаждат разтревожени да ме питат ще летят ли изобщо самалети 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 мосова терапия. Четящите и знаещи хора едва ли се нуждаят от нея.8 ежедневната ни представа Земята е нешо 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 пластична субстанция, която геофизиците наричат мантия (обвивка). В центъра е желязно-никелово ядро с радиус цели 3470 км, но и то в по-голямата си външната си част е течно.Как знаем всичко това след като никой не е бил там, нито пък е възмажно 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 температурата на средите, през които са преминали.Като кажем "крежко" 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. вместо средна дебелина от 35 км, земната кора би трябвало да е 175 км. Впрочем под България тя варира от около 35 до 50 км. При това Земята представлява натрошено яйце, с черупки отделните континентални плочи. Границата между две такива черупки минава по цялата дължина на Атлантическия океан и през Исландия. Плочите свободно "плуват" върху мантията. За щастие материалът, от който е съставена манията, е с огромен вискозитет, т. е. тече бавно и върху него се плува бавно. Много по-бавно от захаросан мед, континентите се движат със скорост от около сантиметър на година! Там, където плочите се движат една към друга, се образуват планини и натрупаната енергия се освобождава чрез земетресения. Хималаите, например, са се формирали от сблъсъка на Индийския субконтинет Азия. Там, където плочите се раздалечават, липсвацият материал във вид на гореща магма се попълва от вътрешността и... пак предизвиква земетресения. Само на отделни места, магмата успява да пробие и предизвиква вулкани. Те могат да се образуват и при насрещно движещи се плочи, ако вместо да се сблъскат, едната се приплъзне под другата. В двата случая наличието на вода улеснява изригването. Началното си загряване планетата ни е получила от енергията на всички парчета, от които се е "слепила". Но това не стига. Лорд Келвин още в средата на XIX век е пресметнал, че Земята без допълнителен източник на енергия би изстинала за някакви си 60 милиона години. Този източник е друг "виновник" за масова катастрофическа тревожност - радиоактивността. Определени елементи са нестабилни, чрез ядрена реакции се превръщат в стабилни и отделят огромна на единична реакция енергия. Т. е. ние живеем върху много бавно действащ естествен ядрен реактор. Макар и не определяща за климата, без топлината от този реактор, Земята щеше да е значително по-негостоприемно място за живеене. Още по лошо, без тектоничните процеси, цялата Земята отдавна щеше да бъде едно равно място, без океани, плонини и континенти. За да извърши това, на ерозията са й необходими само няколко милиона години. Т. е. животът, поне във формите, които познаваме, а в това число и самите ние, нямаше да го има.Изглежда ли сега Ейяфялойскул по-поносимо бедствие? Всышност шената, както и жертвите при природните бедствия са в пряка зависимост от нивото на познания и технологии в обществото. Земетресението, убило толкова хора в Хаити, може изобщо да не отнеме човешки живот в Япония. При същата, че и по-висока степен по Рихтер. Целият свят се възхищава тия дни на умните и организирани исландци, които без да искат помощи, работят, за да ликвидират последиците на своя земя. На фона на техния пример хаосът, предизвикан в Европа със забраната на полети, породи яростни критики. Но око бъдем откровени, при недостатъчната информация за концентрацията и разпространението на вулканичния прах, европейските регулатори правилно не рискуваха живота на екипажи и пасажери. Гледам със завист прекрасния път, който води към почти безлодния район на Ейяфялойокул /виж снимката/ и ми се приисква да можеше да се търгува не само с вредни емисии, а и с бедствия. Аз бих предложил да вземем вулкана, а в замяна исландците да вземат нация образователен "елит", който да им отвори куп университети с прекрасни условия за обучение по всичко... без наука. И тъй като науката не е необходима на новите "университети", важна задача на нашия елит в Исландия ще е да извърши и такава реформа в тяхното начално и средно образование, че исландчетата да не се мъчат повече с природни науки и математика. Е, в бъдеще и том ще усвояват така пътните технологии, че след първата година по пътищата им ще се отварят кратери, сравними с Ейяфялайокул. Но пък ще имат блестящи бизнес администратори, масови комуникатори и прочие търсени от пазара професии.Авторът е физик, старши изследовател 1 Column: 1147 @ Plain Text 1 ⊙ v Tab Size: 4 1 -



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

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

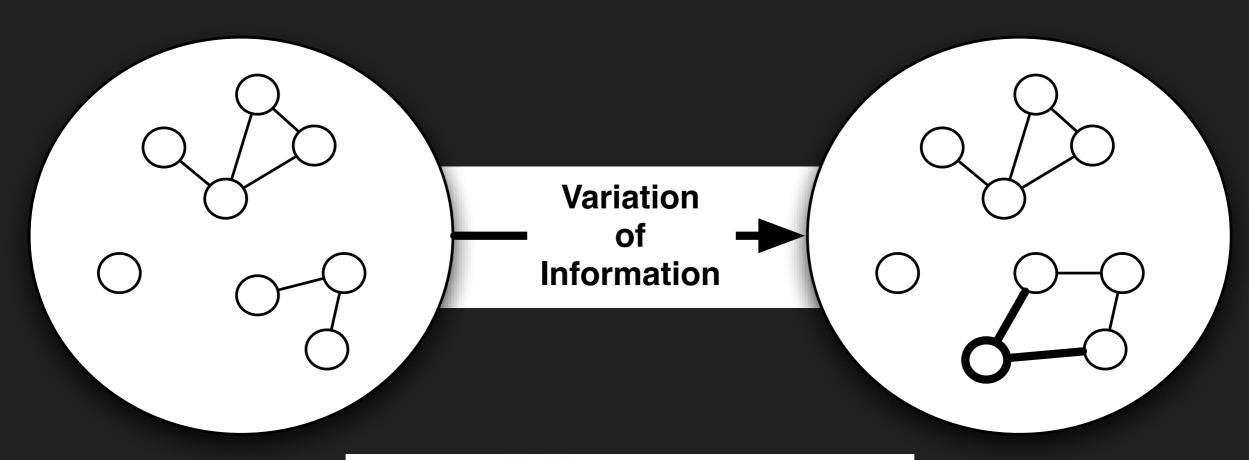


The life expectancy of a node to graph G is given by the time to live (TTL) of this node

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 decremened



Variation of information (Meilã, 2007)

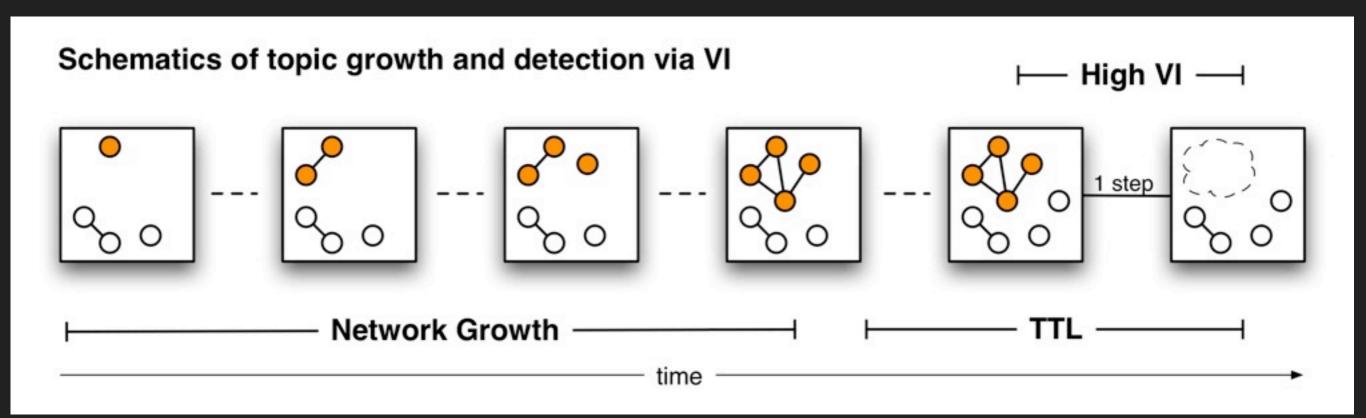


$$I(C,C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P(k')}$$

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



Topic detection



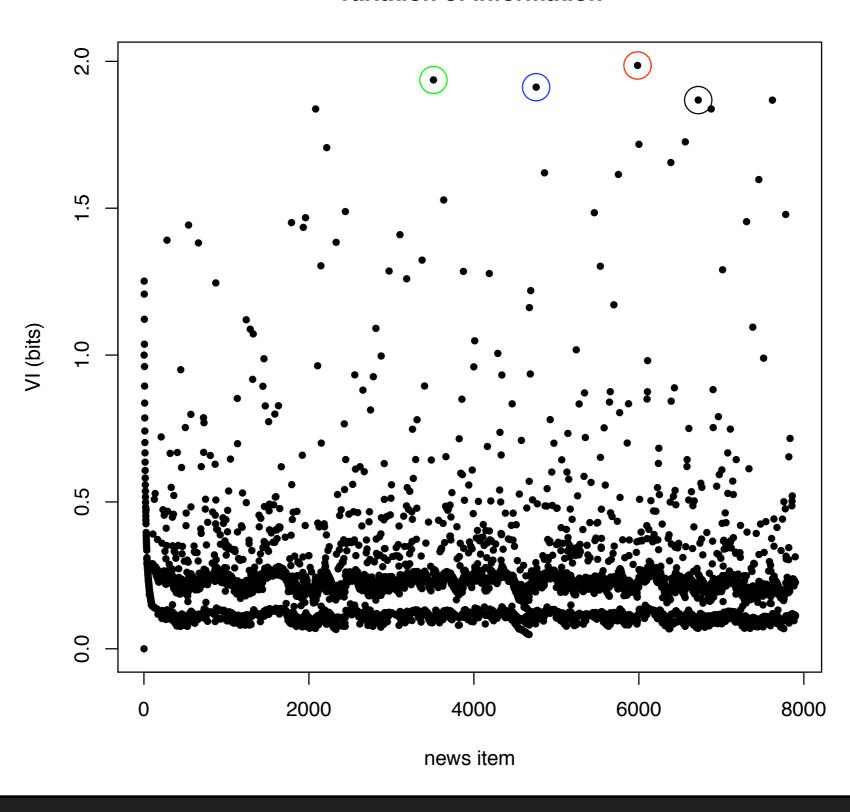
3. Results

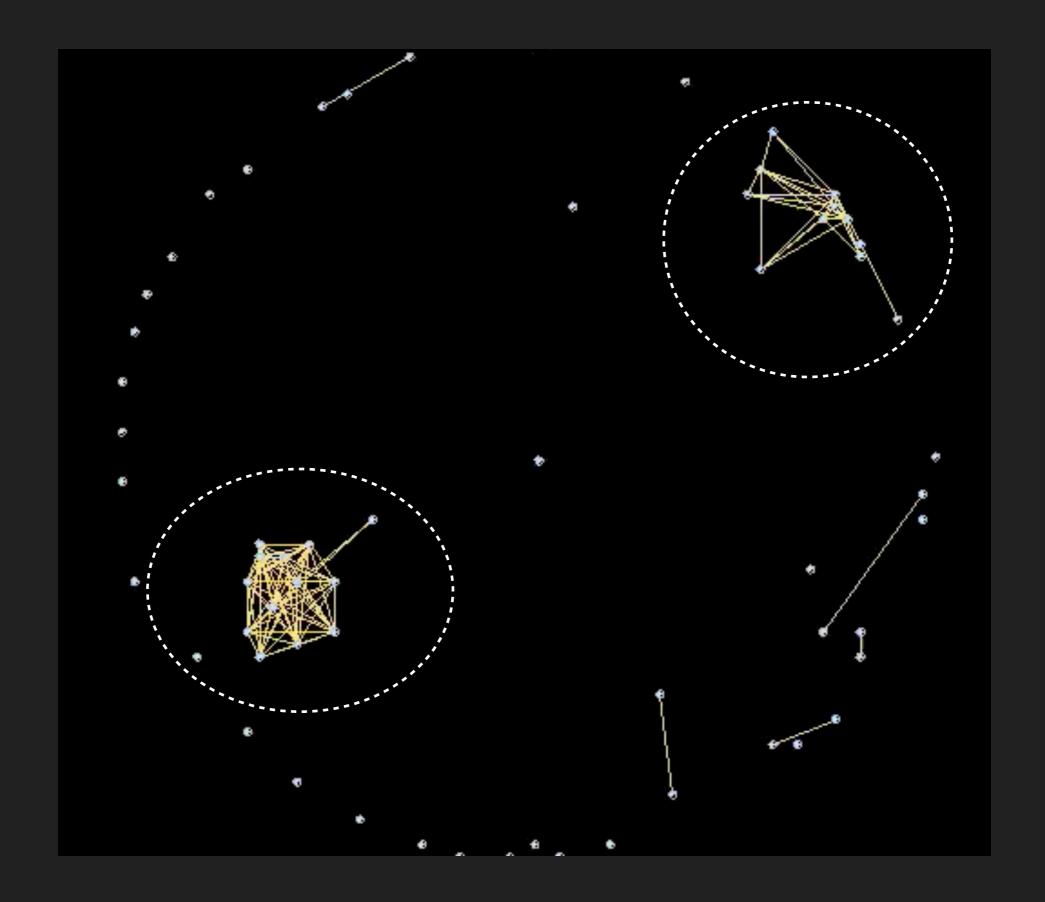


Example: 7928 news items collected from the *Público* newspaper from Nov. I I, 2009 to Jan. 25, 2010



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 newspappers
- 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



- 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?

- 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?