8th September 2016

2nd International Keystone Conference

Cluj-Napoca, Romania

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

UNIMORE

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Universidad Zaragoza 1542

keystone

IC1302

Topic Detection in multi-channel Italian newspapers

Laura Po[1], Federica Rollo[1] and Raquel Trillo-Lado[2]

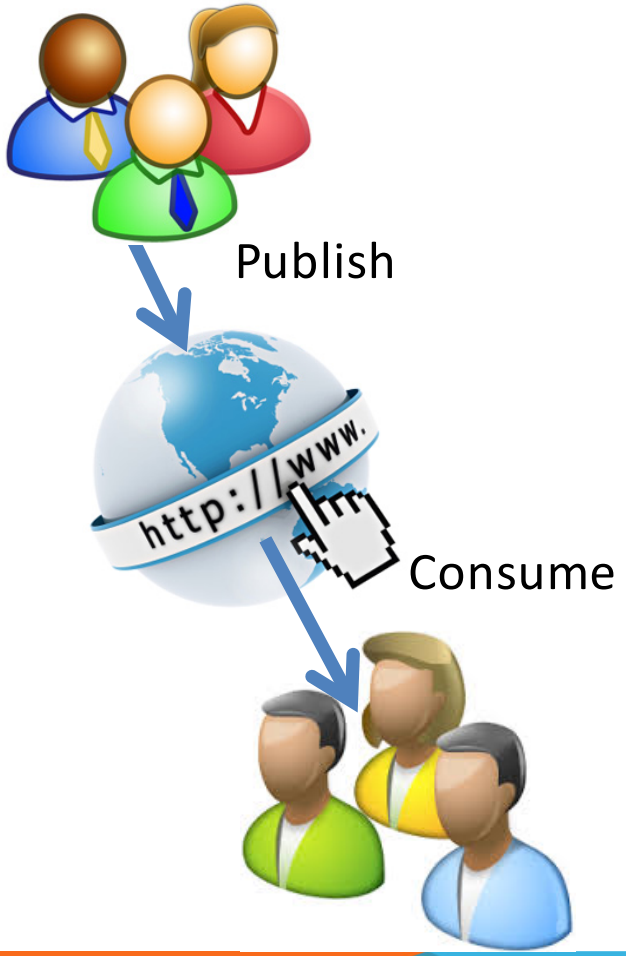
[1]University of Modena and Reggio Emilia (Italy) [2]University of Zaragoza (Spain)

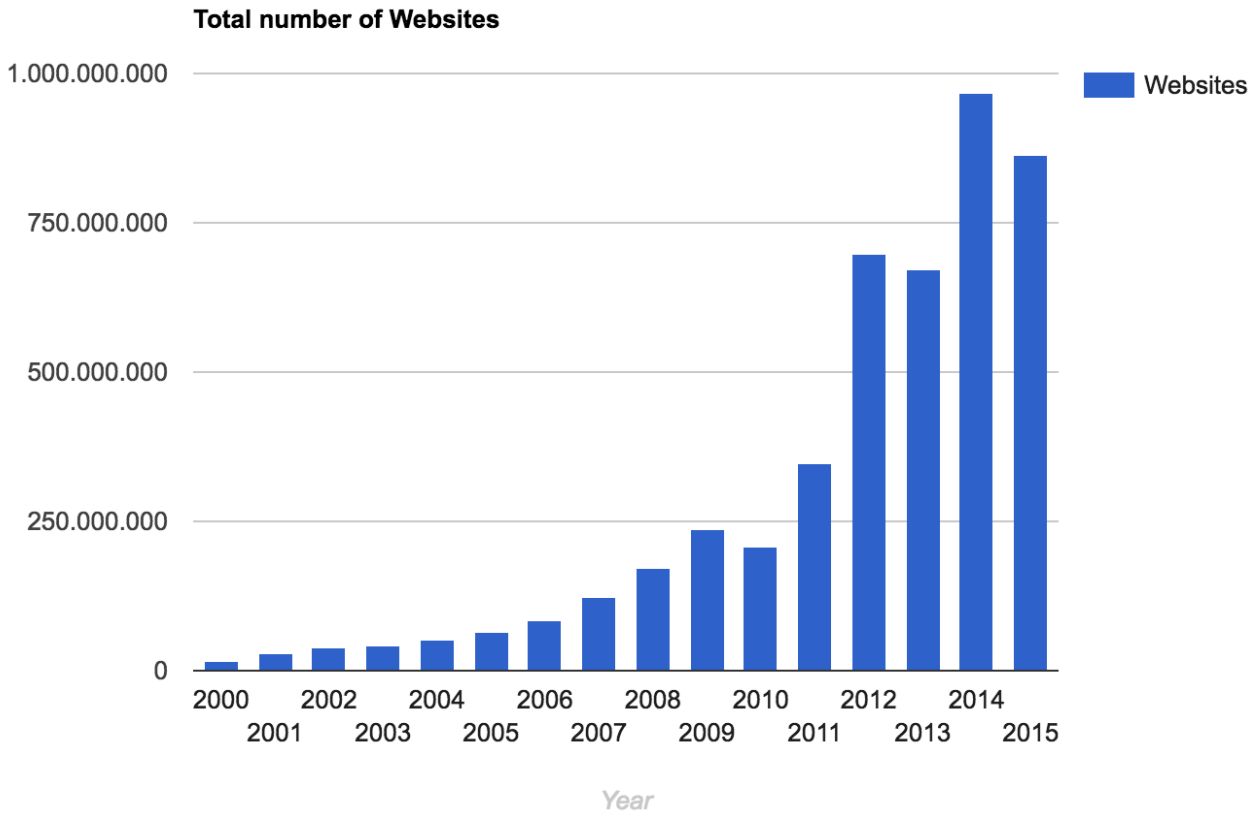
Outline

- **Introduction**
  - **Context:** How ICT has change it the way we consume/share information
  - **Goal:** Clustering content from different channels according to its topic
  - **Motivation:** Why is Topic Detection important?
- **Proposal**
  - **Background:** KeyGraph Algorithm
  - **Adaptation of KeyGraph to the context:** multichannel news vs text documents
- **Evaluation**
  - **Scenario 1:** Correlations among news from one newspaper
  - **Scenario 2:** Correlations among news from a set of newspapers
- **Conclusions and Future Work**

Introduction

- **Context:** How ICT has changed the way information/data are consumed/shared





[Extracted from: http://www.internetlivestats.com/]

Introduction

- **Context:** How ICT has changed the way information/data are shared/consumed





- **Web Editions vs Printed Editions**
  - It's immediately accessible from any place
  - It requires less time to be updated
  - It's cheaper

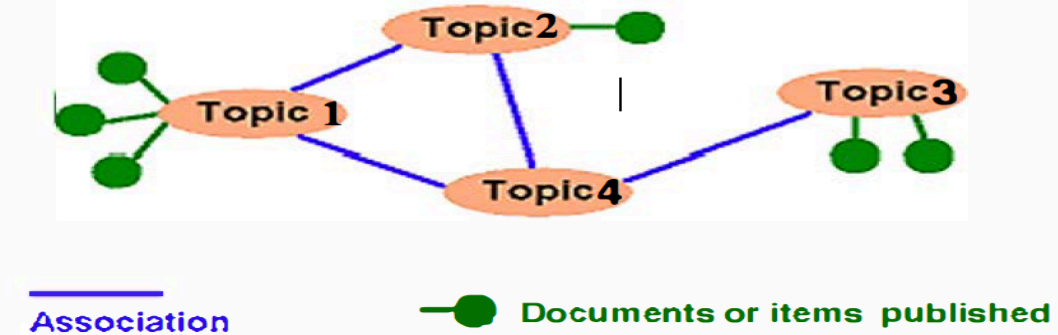
Introduction

- **Context:** How ICT has changed the way information/data are shared/consumed
  - **Boom of Social Web:**



Introduction

- **Goal:** Analyzing and clustering the content (published by an user/company or a sector) according to their topics



- **Traditional approaches**
  - **Single Channel**

Our approach

Multiple Channel

- Different writing styles depending on the channels
- Producers can have different purposes for the different channels

Introduction

- **Motivation: Why is Topic Detection important?**

  - A **Topic Map** is a **comprehensive overview** of the information published by a source.

  - **Topic Maps** can be used to determine if there **exist correlations among the items** published by **different sources** (users, companies, etc.) **on different channels.**

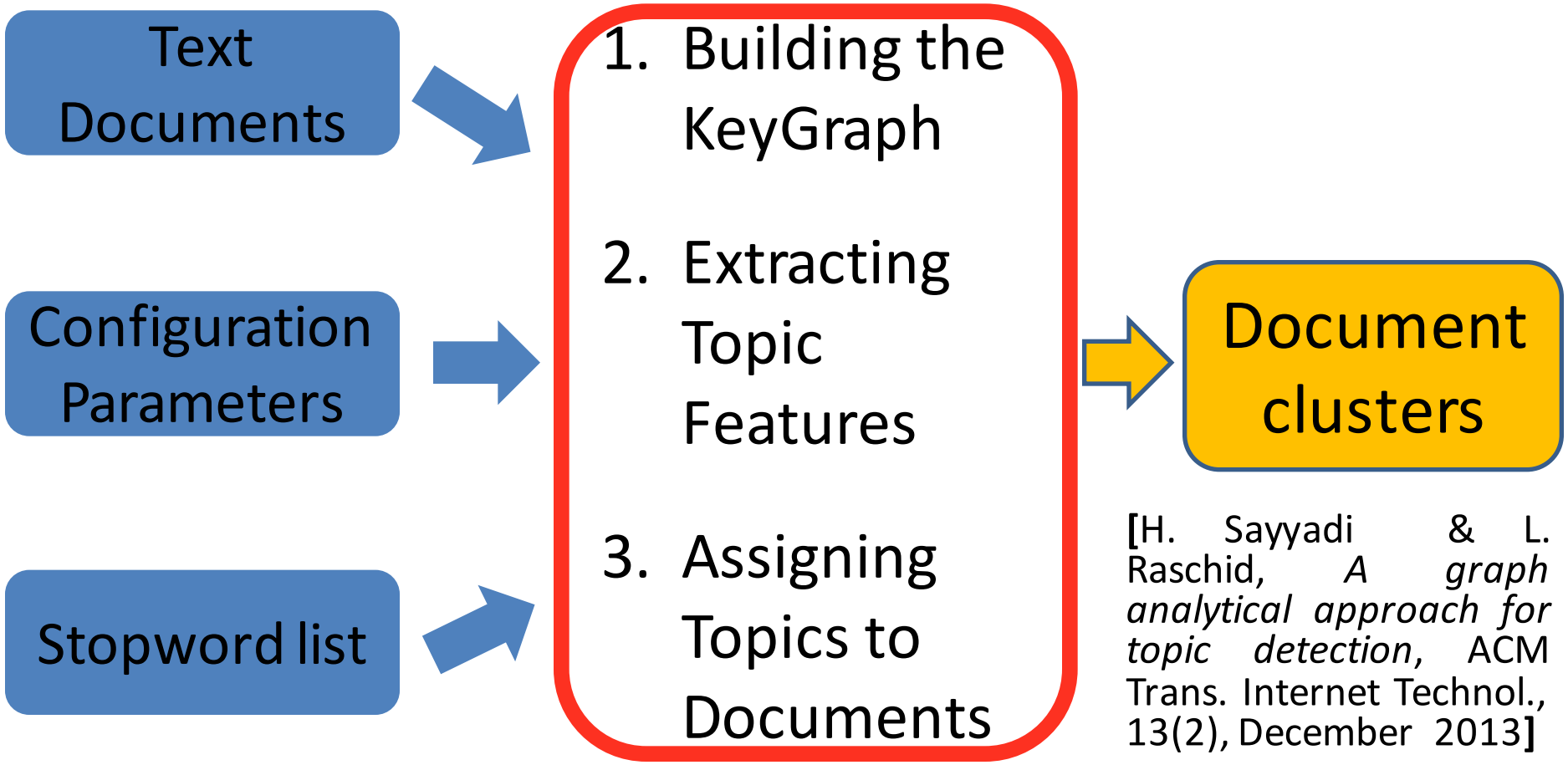
  - The **topics** can be used **to characterized the items of collections** in order to suggest users new items **(**i.e., they are used on **Recommendation Systems)**

Outline

- **Introduction**
  - **Context:** How ICT has change it the way we consume/share information
  - **Goal:** Clustering content from different channels according to its topic
  - **Motivation:** Why is Topic Detection important?
- **Proposal**
  - **Background:** KeyGraph Algorithm
  - **Adaptation of KeyGraph to the context:** multichannel news vs text documents
- **Evaluation**
  - **Scenario 1:** Correlations among news from one newspaper
  - **Scenario 2:** Correlations among news from a set of newspapers
- **Conclusions and Future Work**

Proposal

- **Background: KeyGraph Algorithm**



[H. Sayyadi & L. Raschid, *A graph analytical approach for topic detection*, ACM Trans. Internet Technol., 13(2), December 2013]

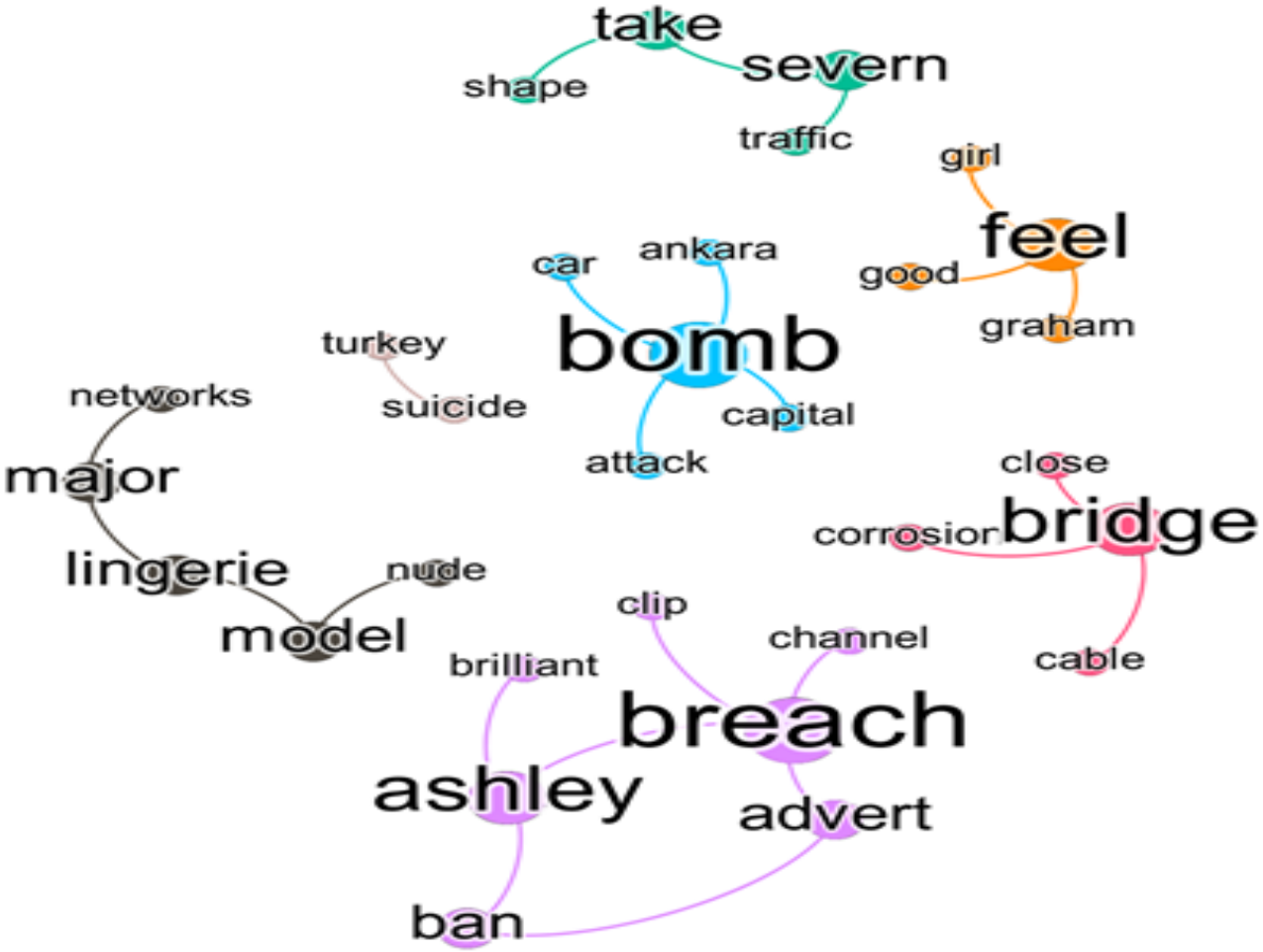
Proposal: Background

- **Phase 1: Building the KeyGraph**

ID	CHANNEL	AUTHOR	BODY	DATE
0	Twitter	The Mirror	At least 27 killed and 75 wounded in 'suicide bombing' in Turkish capital #Ankara.	13-03-2016
1	Facebook	The Independent	Ashley Graham lingerie advert 'banned by major TV networks'	13-03-2016
2	Site	The Sun	Plus-size models have nude advert REJECTED by stuffy TV channels for being 'indecent'. Girls in lingerie told they breach rules with feel-good clip. #Ashley #Graham	13-03-2016
3	Twitter	Broadway Vans	One lane of Severn Bridge to close until #October for cable inspection. #traffic	08-03-2016

Proposal: Background

- **Phase 2: Extracting Topic Features**



Proposal: Background

- **Phase 3: Assigning Topics to Documents**

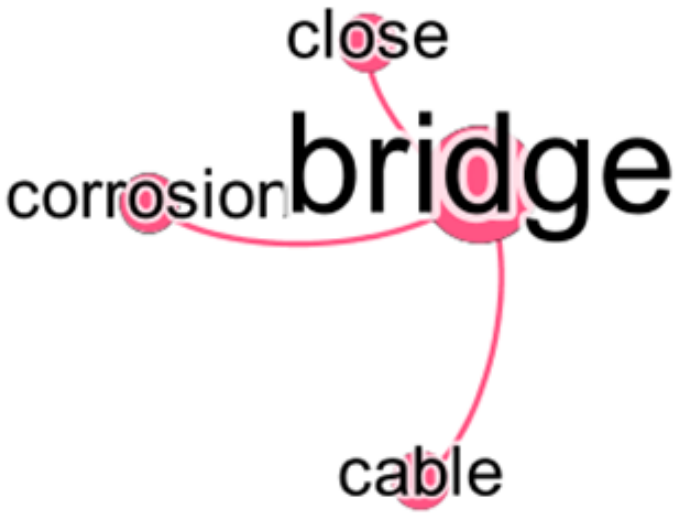
$$V = \sqrt{\sum_{n=0}^{k} TF_{\downarrow} IDF_n^2}$$, n is the number of nodes considered

$$cosine_{\downarrow} similarity = \frac{\sum_{n=0}^{k} (TF_{\downarrow} IDF_{n_{comm}} * TF_{\downarrow} IDF_{n_{doc}})}{V_{document} * V_{community}}$$

n common nodes between the community and the document considered

Proposal: Background

- **Phase3: Assigning Topics to Documents**



One lane of Severn Bridge to close until #October for cable inspection. #traffic



Old Severn bridge cables to be inspected for corrosion. One lane closed.



Highway closed for 24 hours: traffic haywire.



Proposal: Adaptation

- **Adaptation of KeyGraph to the context: multichannel vs text documents**

KEYGRAPH' VERSION 1

- Deletion of the author names and mentions 
- Extraction and analysis of hashtags 
- Different Configuration Parameters for each channel   

KEYGRAPH' VERSION 2

- Creation of edges (and nodes if needed) between the keywords of pair of documents with at least one common link

KEYGRAPH' VERSION 3

- In the third phase, guarantee that items that have a common link appear in the same clusters

Outline

- **Introduction**
  - **Context:** How ICT has change it the way we consume/share information
  - **Goal:** Clustering content from different channels according to its topic
  - **Motivation:** Why is Topic Detection important?
- **Proposal**
  - **Background:** KeyGraph Algorithm
  - **Adaptation of KeyGraph to the context:** multichannel news vs text documents
- **Evaluation**
  - **Scenario 1:** Correlations among news from one newspaper
  - **Scenario 2:** Correlations among news from a set of newspapers
- **Conclusions and Future Work**

Evaluation

- **Baseline: Clustering based on links (Link-Clustering)**
  - **If an item/document is in a cluster, then it has at least an URL Link in common with at least one other item/document in the same cluster.**

- **Two different scenarios:**
  - **News** coming from the **channels** (Twitter, Facebook and Website) of a **single source:**





  - **Documents** coming from the **channels** of a **collection of sources: 20 most popular newspapers in Italy**

keystone

Evaluation

Newspaper	Paper editions	Digital editions	Total circulation
Corriere della Sera	368 981	95 447	464 428
Repubblica (La)	323 525	58 709	382 234
Sole 24 Ore (Il)	200 155	115 366	315 521
Gazzetta dello Sport (La)	203 516	21 042	224 558
Stampa (La)	214 461	7 198	221 659
Messaggero (Il)	137 678	4 510	142 188
QN – Il Resto del Carlino	122 513	1 234	123 747
Corriere dello Sport - Stadio	121 128	1 272	122 400
Giornale (Il)	103 658	2 115	105 773
Avvenire (L')	103 985	1 578	105 563
QN – La Nazione	98 812	1 094	99 906
Tuttosport	94 970	818	95 788
Libero	75 301	886	76 187
Italia Oggi	54 166	18 157	72 323
Gazzettino (Il)	66 163	4 276	70 439
Fatto Quotidiano (Il)	50 763	13 621	64 384
Secolo XIX (Il)	57 068	1 208	58 276
Tirreno (Il)	56 639	1 539	58 178
Mattino (Il)	50 946	2 429	53 375
QN – Il Giorno	50 597	232	50 829



Evaluation

- **Scenario 1:** *La Repubblica*

Documents/entries : 2400  18-22 January 2016

	Link Cluster (Baseline)	KeyGraph		
		First version	Second version	Third version
DOCUMENT CLASSIFIED	112	404	290	294
DOCUMENT CLUSTERS	51	82	77	69

Evaluation

- **Scenario 1:** *La Repubblica*

Documents/entries : 2400

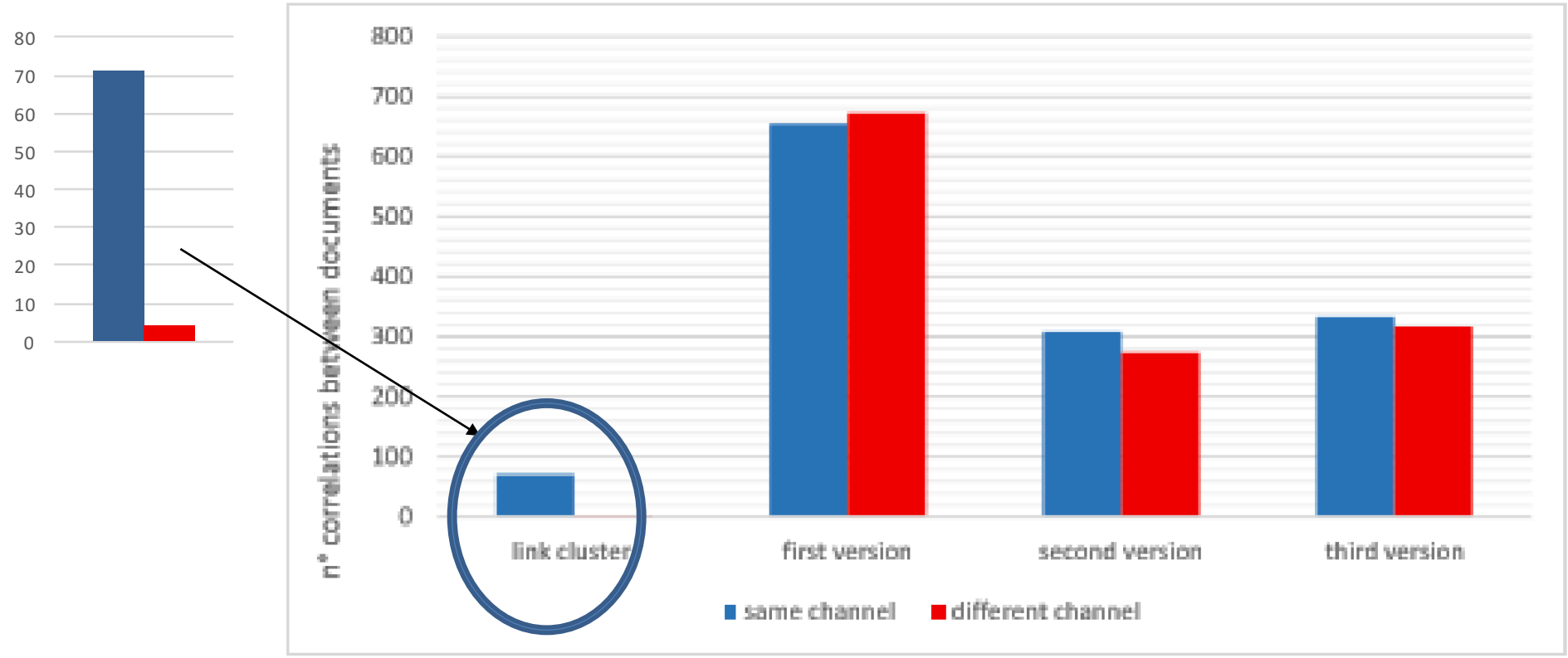


18-22 January 2016

	First version	Second version	Third version
Accuracy	81%	85%	87%
Precision	60%	75%	73%
Recall	65%	69%	73%
Time (seconds)	2.28	3.12	3.36

Evaluation

- **Scenario 1: Correlations among news** from La Repubblica **inside the clusters**



- **Scenario 2:** set of the 20 most popular newspapers in Italy

Documents/entries : 21457



18-22 January 2016

	Link cluster	KeyGraph		
		First version	Second version	Third version
DOCUMENT CLASSIFIED	3233	2777	1025	1074
DOCUMENT CLUSTERS		514	265	261

Evaluation

- **Context 2:** set of the 20 most popular newspapers in Italy

Documents/entries : 21457

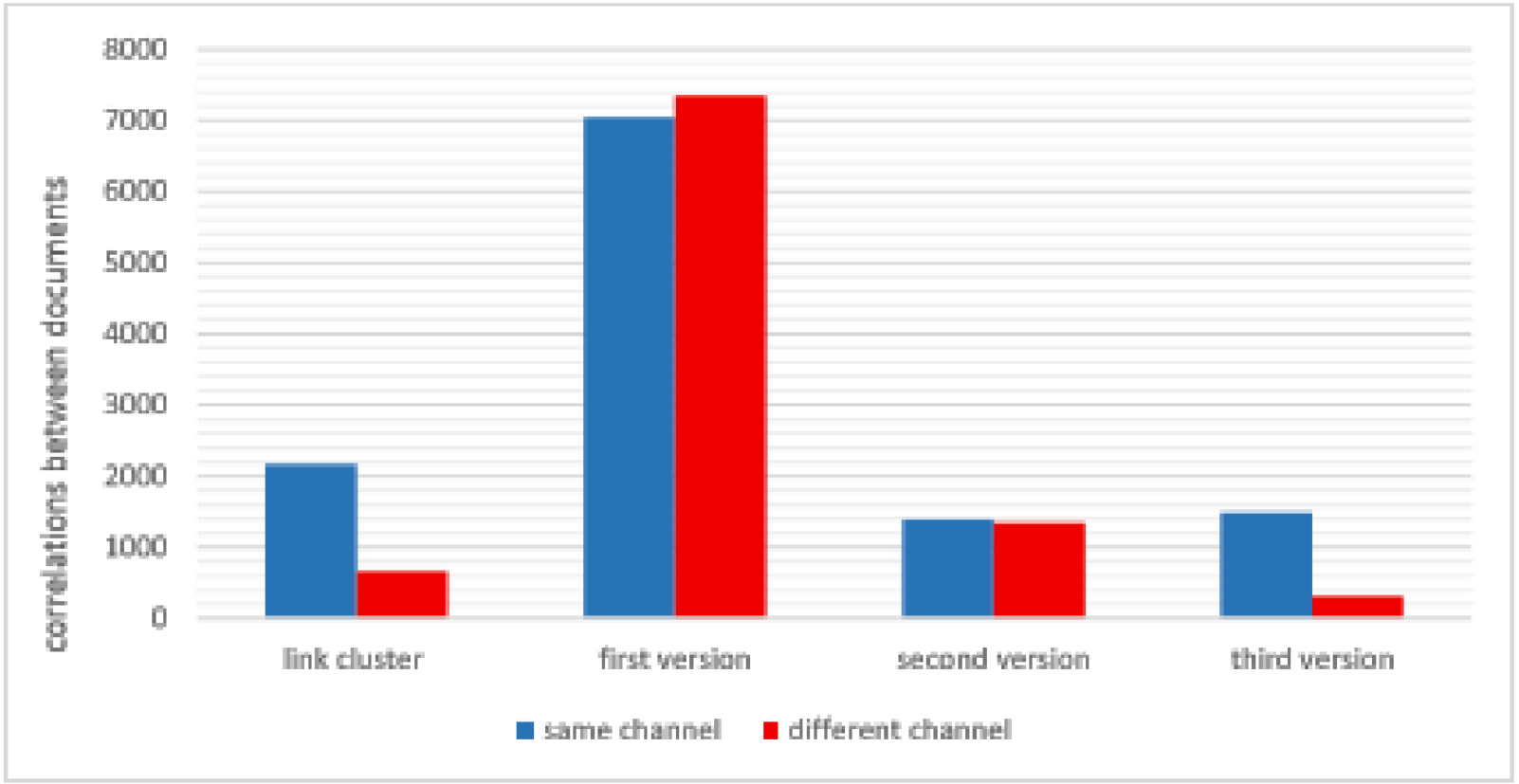


18-22 January 2016

	First version	Second version	Third version
Accuracy	83%	88%	90%
Precision	75%	85%	86%
Recall	75%	79%	82%

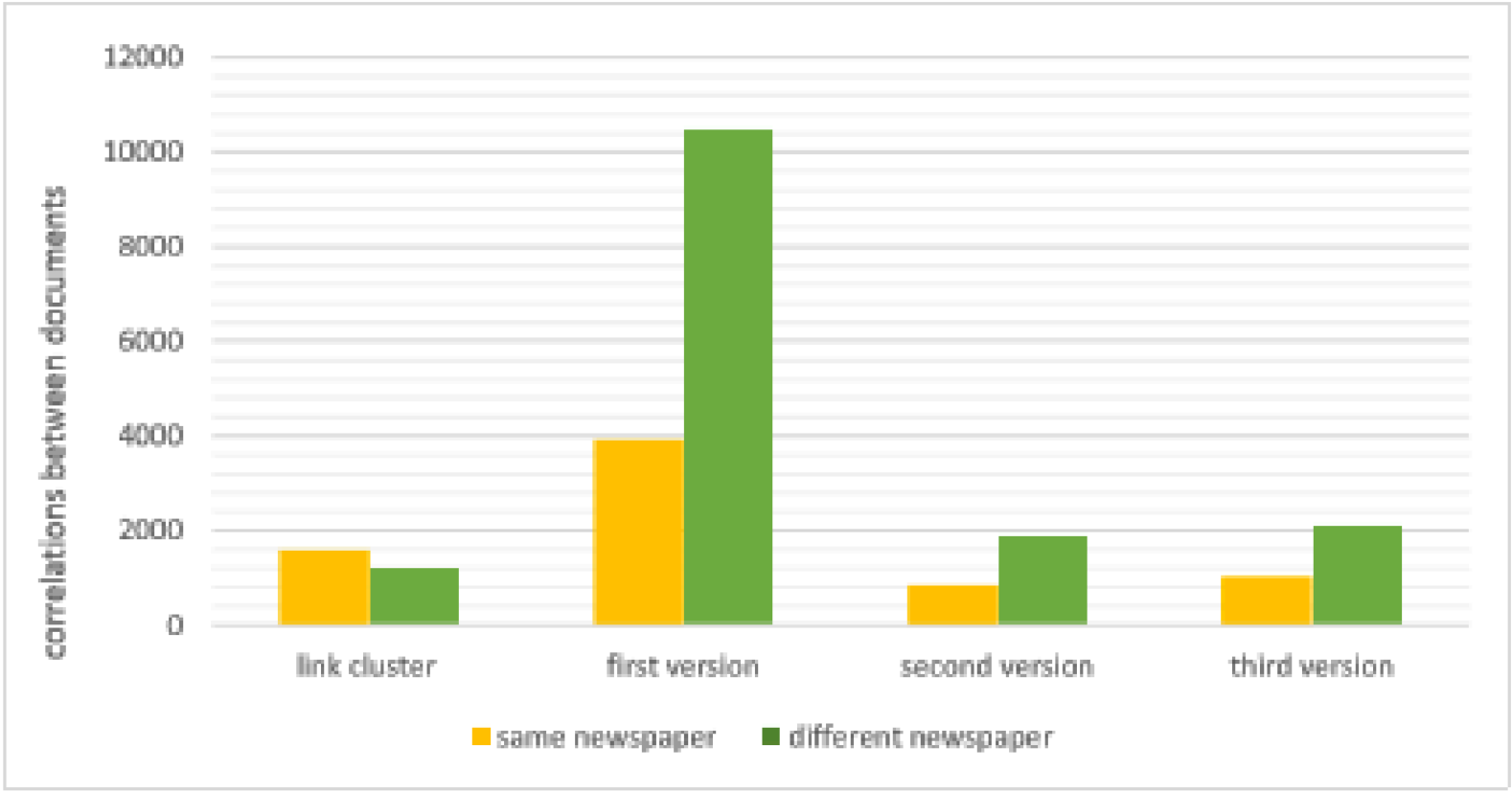
Evaluation

- **Scenario 2: Correlations among news from 20 popular Italian newspapers**



Evaluation

- **Scenario 2: Correlations among news from different data-sources (newspapers)**



Outline

- **Introduction**
  - **Context:** How ICT has change it the way we consume/share information
  - **Goal:** Clustering content from different channels according to its topic
  - **Motivation:** Why is Topic Detection important?
- **Proposal**
  - **Background:** KeyGraph Algorithm
  - **Adaptation of KeyGraph to the context:** multichannel news vs text documents
- **Evaluation**
  - **Scenario 1:** Correlations among news from one newspaper
  - **Scenario 2:** Correlations among news from a set of newspapers
- **Conclusions and Future Work**

Conclusions and Future Work

- **Conclusions:**
  - **Adaptation of KeyGraph** algorithm to analyze and cluster items coming from **multiple-channels**
  - Wide set of **experiments** considering the news published by the main Italian **newspapers**
  - A valid approach to **analyze correlations** and **make comparison** among items (news) published on **different channels** and **data sources** (newspapers)

- **Future Work:**
  - Studying the Adaptation of KeyGraph in **different context (benchmarks)**
  - Considering **synonyms and disambiguation techniques** to create the initial graph
  - Analyzing the results by considering **different configuration parameters**

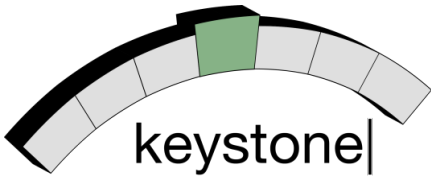
Thank you very much for you attention



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



Universidad
Zaragoza



IC1302

Topic Detection in multi-channel Italian newspapers

Laura Po[1], Federica Rollo[1] and Raquel Trillo-Lado[2]
{laura.po, federica.rollo]@unimore.it[1] and raqueltl@unizar.es[2]