sucinto info

summarization
for clever information access

Graph-based Methods for Multi-document Summarization

Exploring Relationship Maps, Complex Networks and Discourse Information

Rafael Ribaldo, Ademar T. Akabane,
Lucia H. M. Rino, Thiago A. S. Pardo

ICMC USP SÃO CARLOS SCC

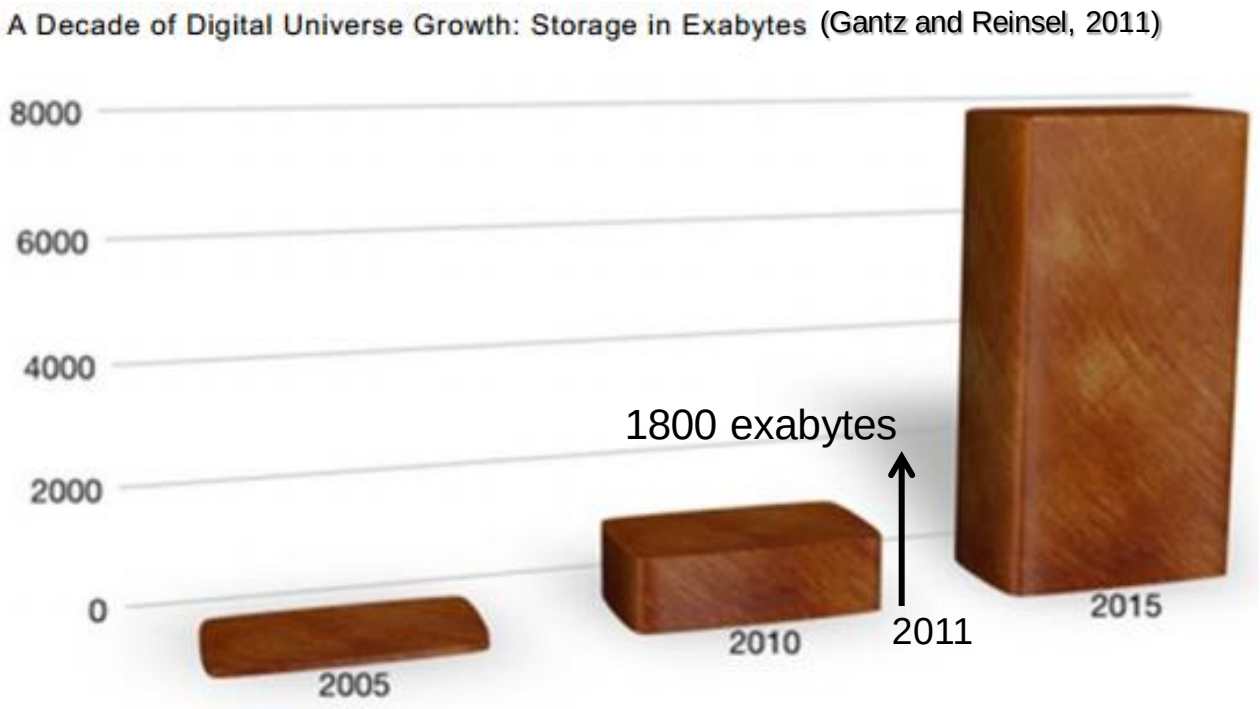
DC Departamento de Computação

NILC

CNPq FAPESP

Multi-document summarization (MDS)

- *Automatic production of a unique summary from a group of texts on the same topic* (Mani, 2001)



Some history for Portuguese

- First works for English in the 90s (McKeown and Radev, 1995)

- For (written) Portuguese
  - Superficial methods
    - GistSumm (Pardo, 2005)
    - Combination of superficial methods (Alves et al., 2007)
  - Deep methods
    - CSTSumm (Castro Jorge and Pardo, 2010)
    - Discourse-based methods for MDS (Cardoso et al., 2011)
  - Machine learning (also using deep knowledge)
    - Discriminative learning (Castro Jorge et al., 2011)
    - Generative learning (Castro Jorge and Pardo, 2011)

- PROPOR 2012
  - SIMBA (Silveira and Branco, 2012)

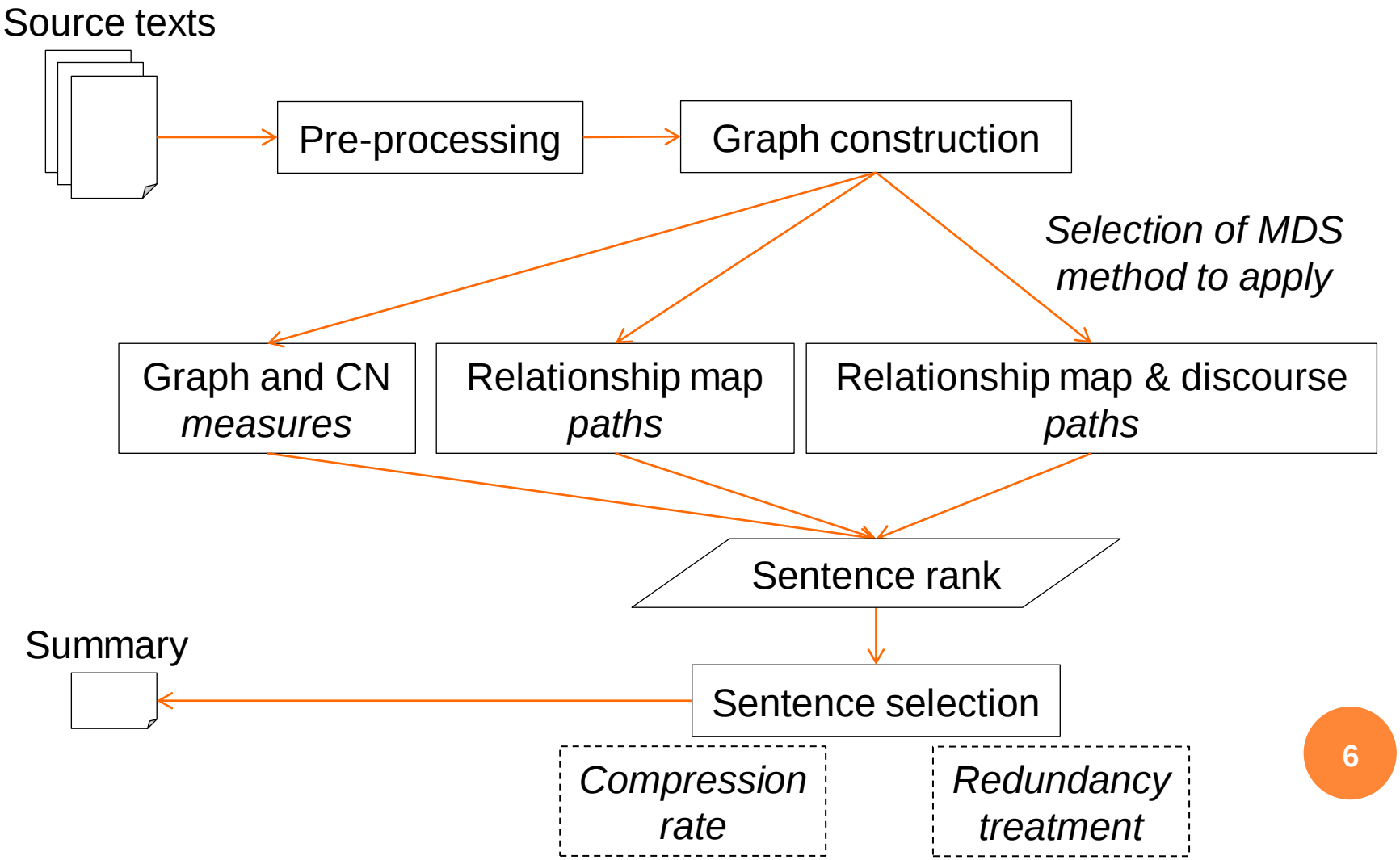
This work

- Investigation of some graph-based methods for content selection in MDS
  - Relationship maps (Salton et al., 1997)
    - Classical approach
  - Graph and complex network measures (Antiqueira et al., 2009)
    - Recent trend
      - *Elegant*, *scalable* and *good approaches* to the problem
      - *Increasing interest* for summarization (Erkan and Radev, 2004; Mihalcea et al., 2005, 2006; Wan, 2008)

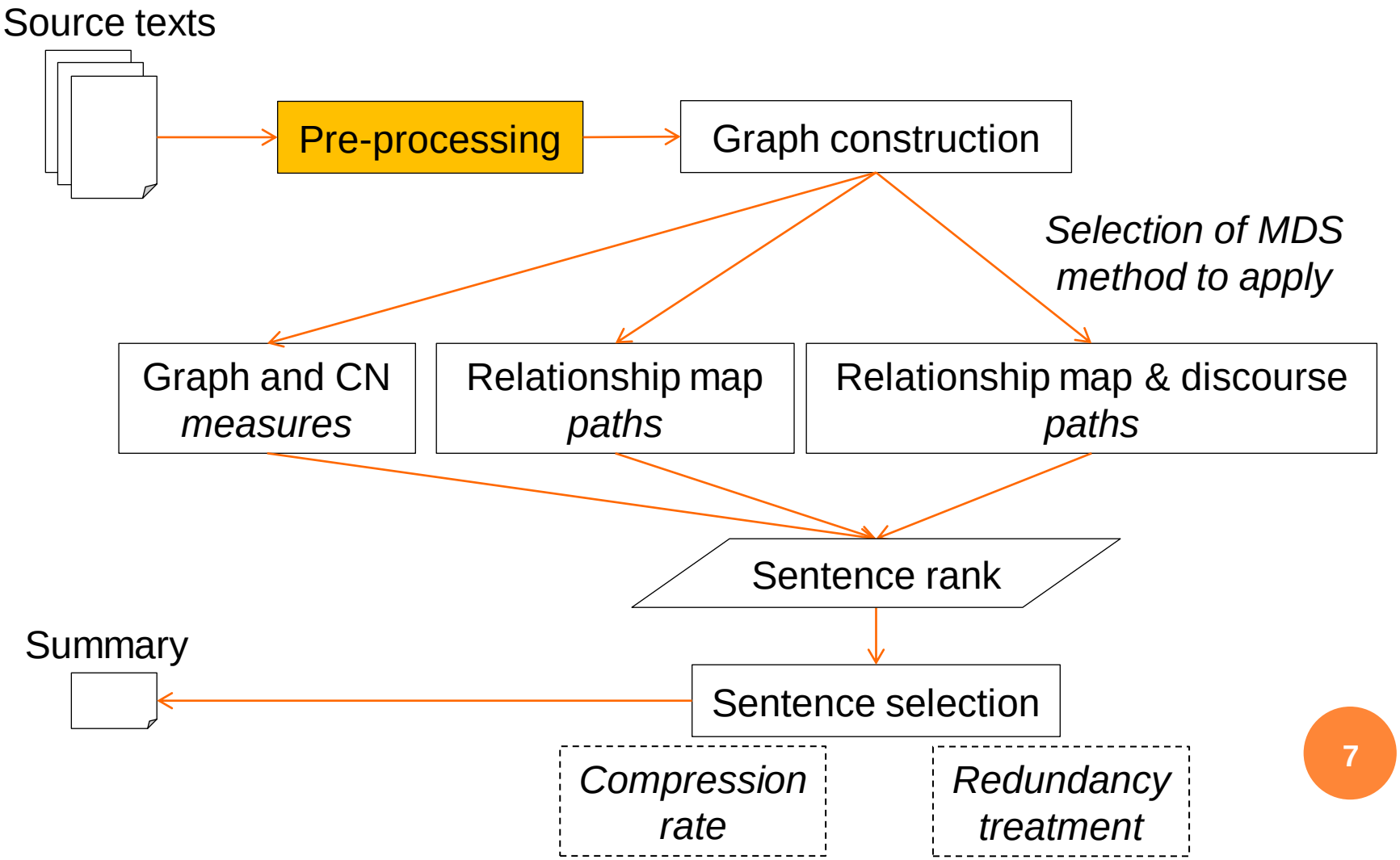
This work

- Investigation of some graph-based methods for content selection in MDS
  - Impact of discourse information in the methods
    - Cross-document Structure Theory – CST (Radev, 2000)
      - Redundancy, information overlap, contradictions, writing style differences, etc.
        - *Heavily used* in current MDS works for Portuguese

METHOD – OVERVIEW



METHOD – OVERVIEW

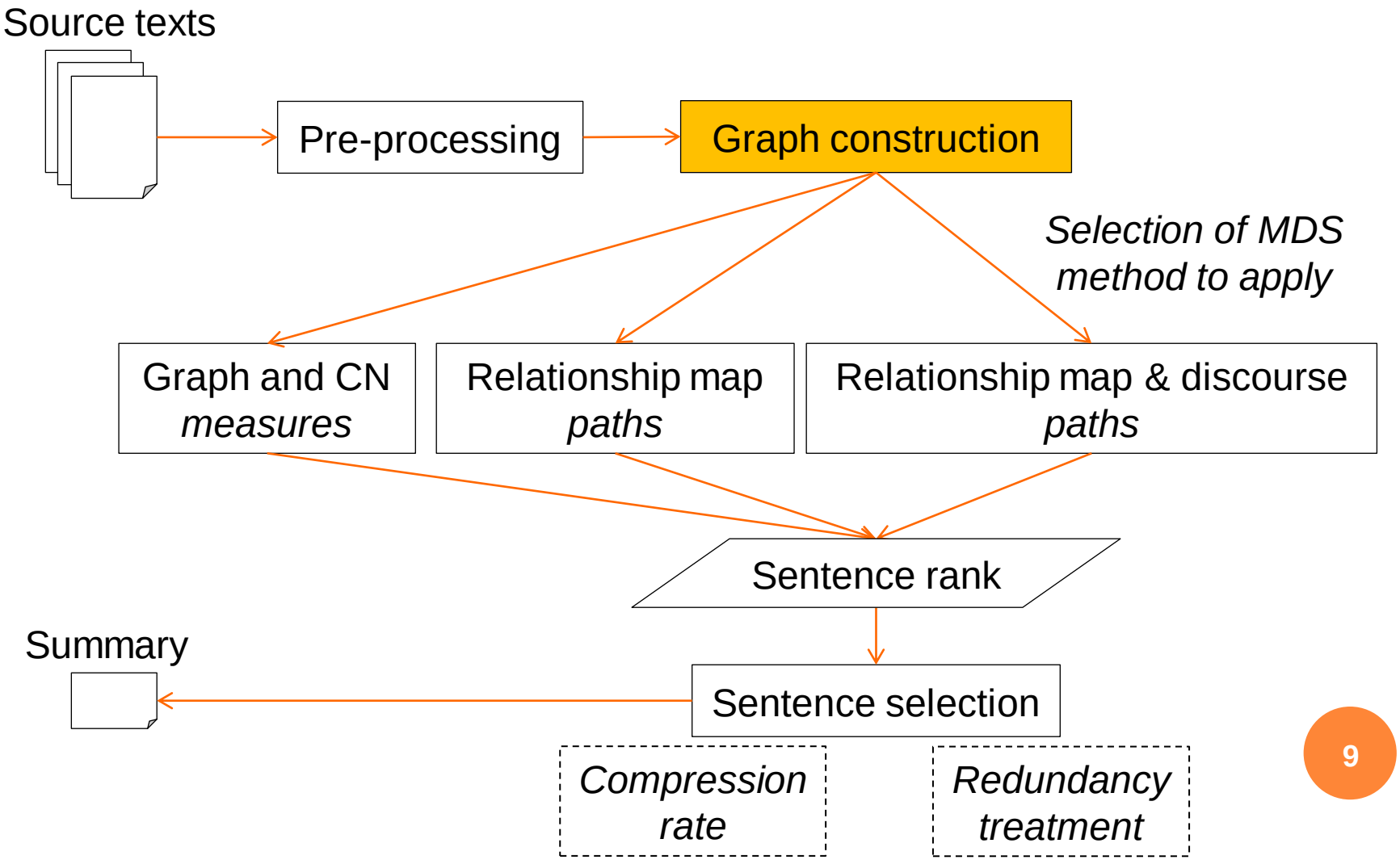


Step by step

1. Pre-processing the source texts

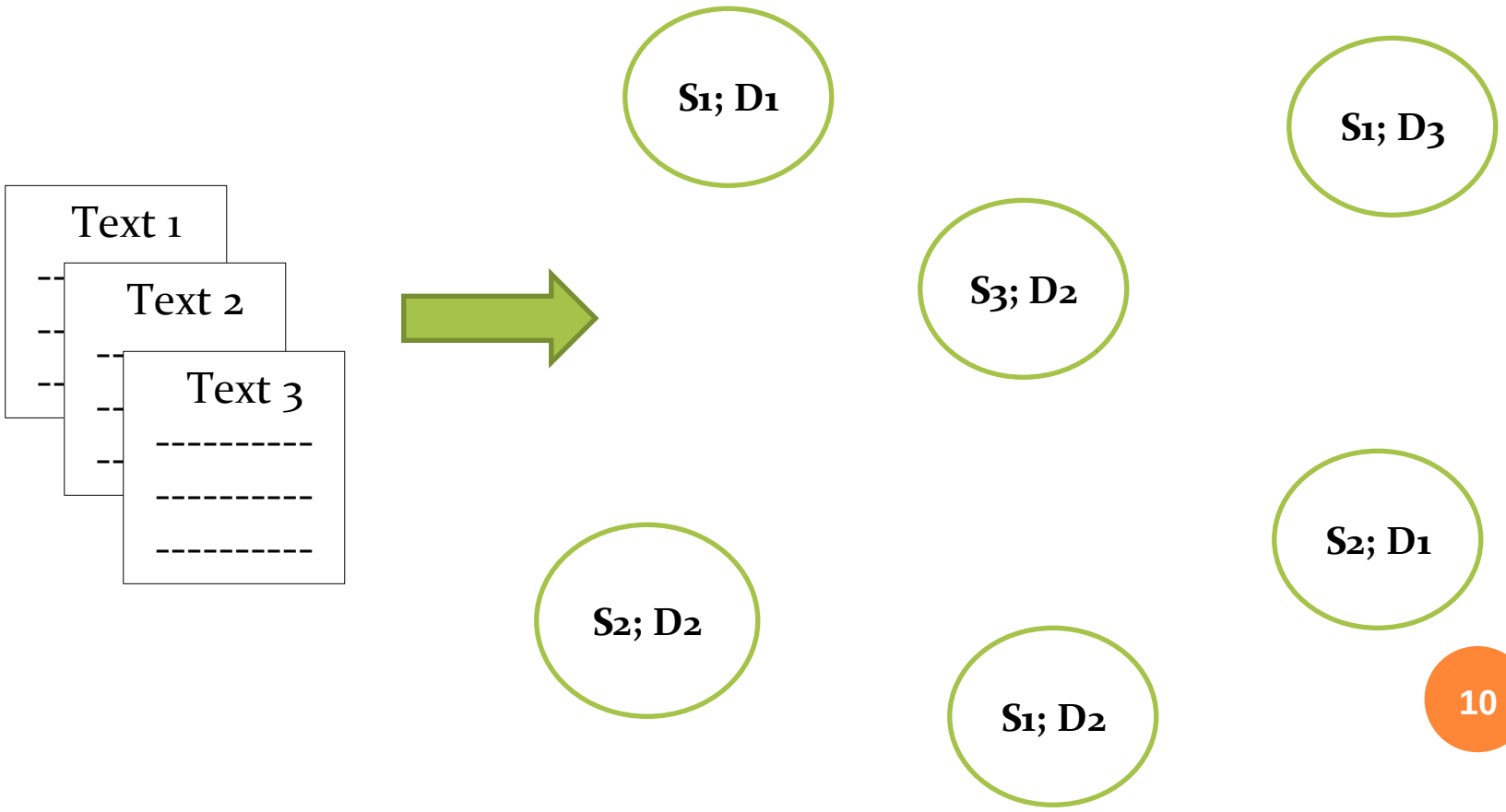
- Tokenization and sentence segmentation
  - SENTER (Pardo, 2006)
- Case folding
- Stopwords removal
- Stemming
  - Snowball Portuguese stemmer

METHOD – OVERVIEW



Step by step

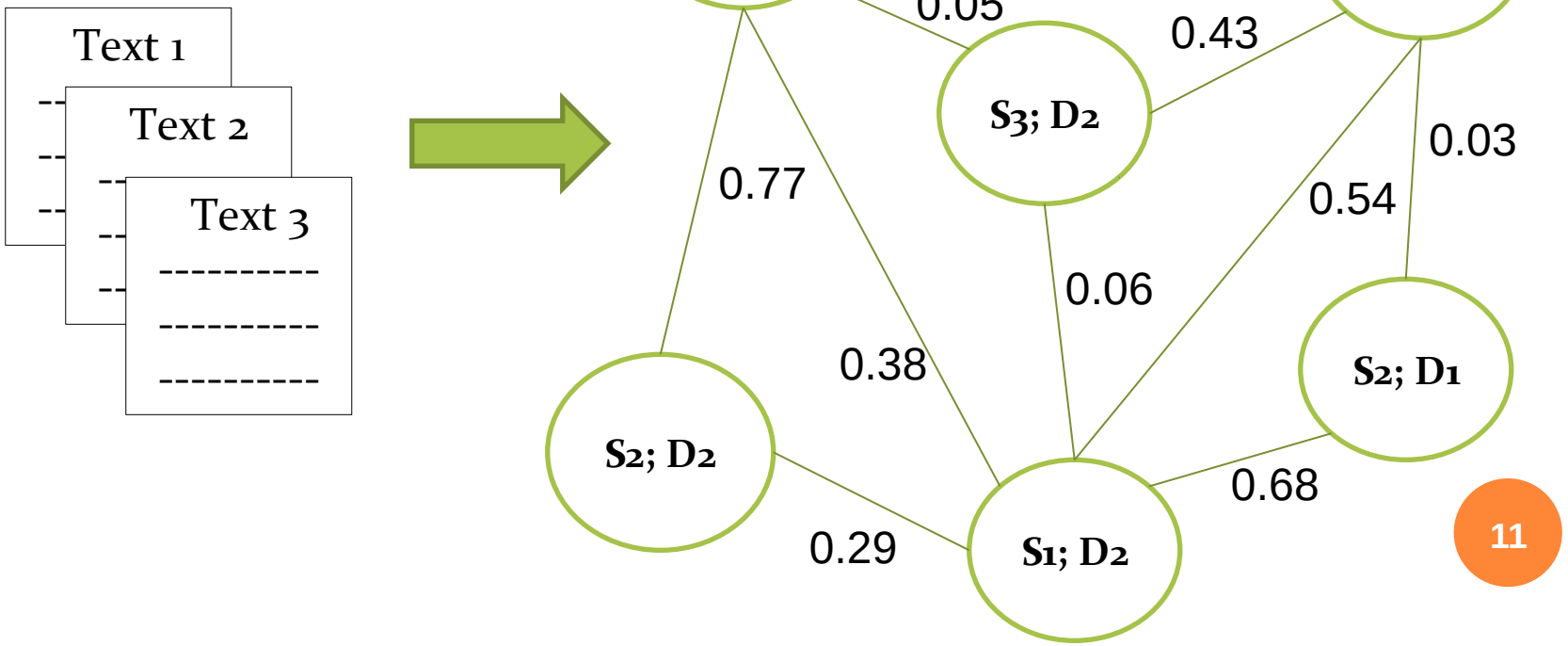
2. Modeling source texts as a graph
   - Sentences as nodes



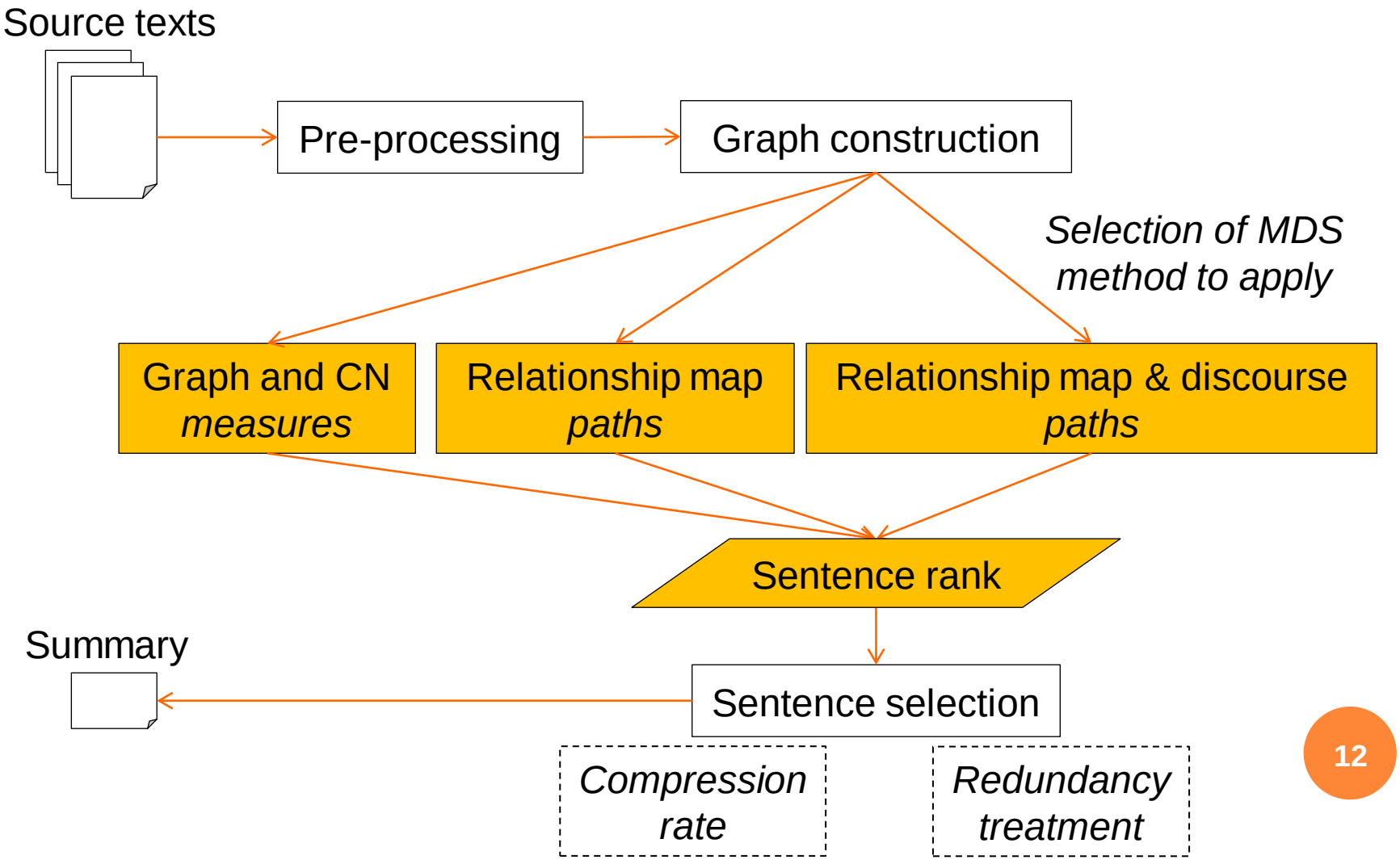
Step by step

2. Modeling source texts as a graph

- Weighted edges
  - Cosine measure (Salton, 1988)



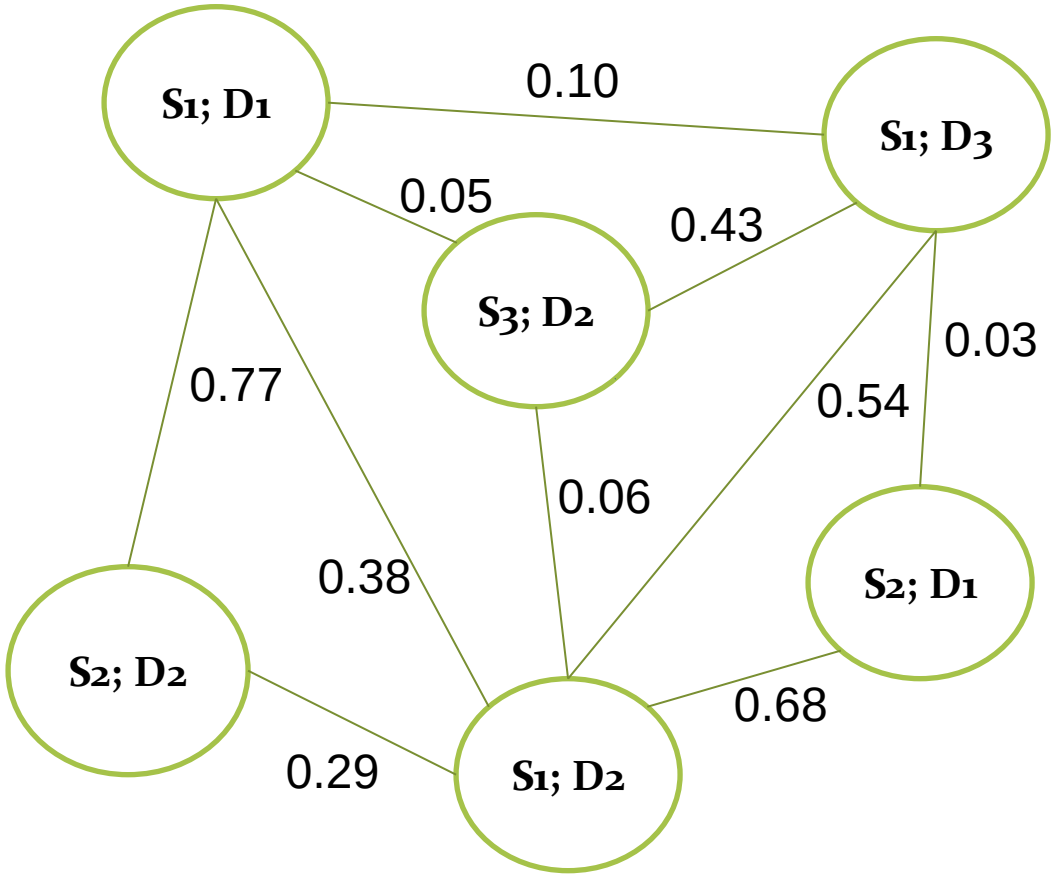
METHOD – OVERVIEW



Step by step

3.1. Graph and complex network measures

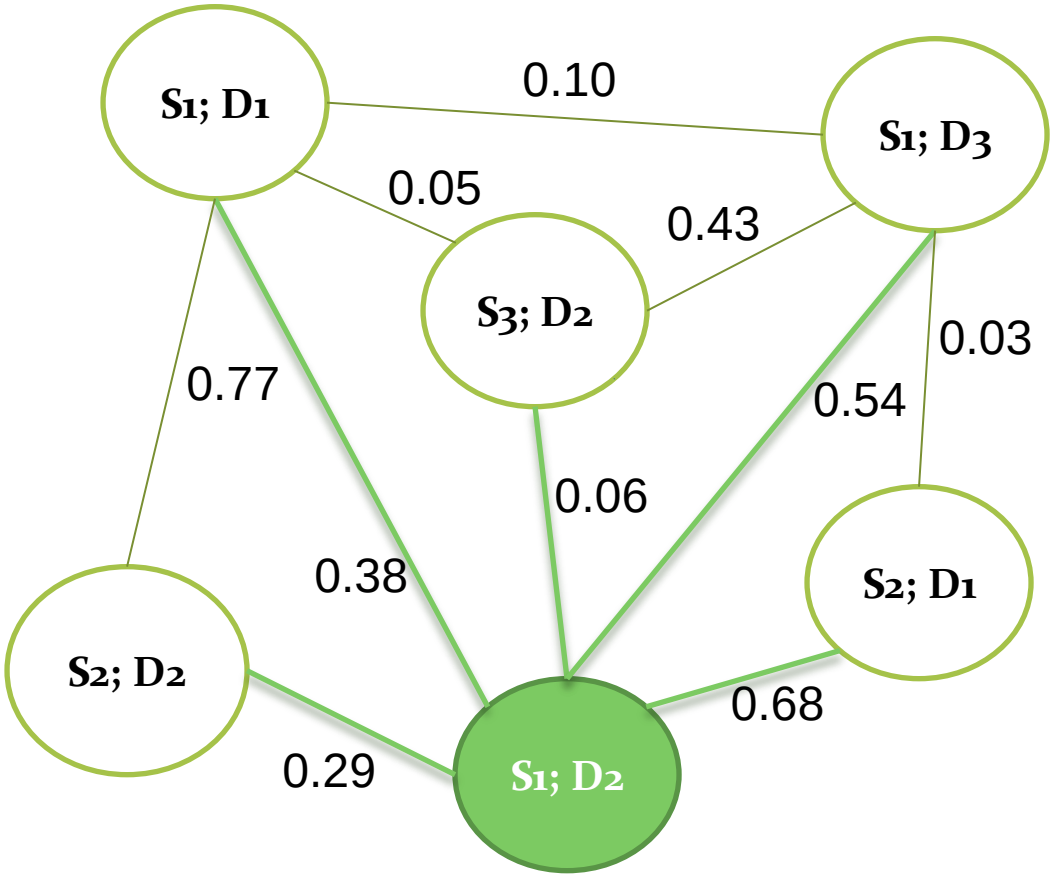
- Degree, avg. shortest path, clustering coefficient



Step by step

3.1. Graph and complex network measures

- Degree, avg. shortest path, clustering coefficient

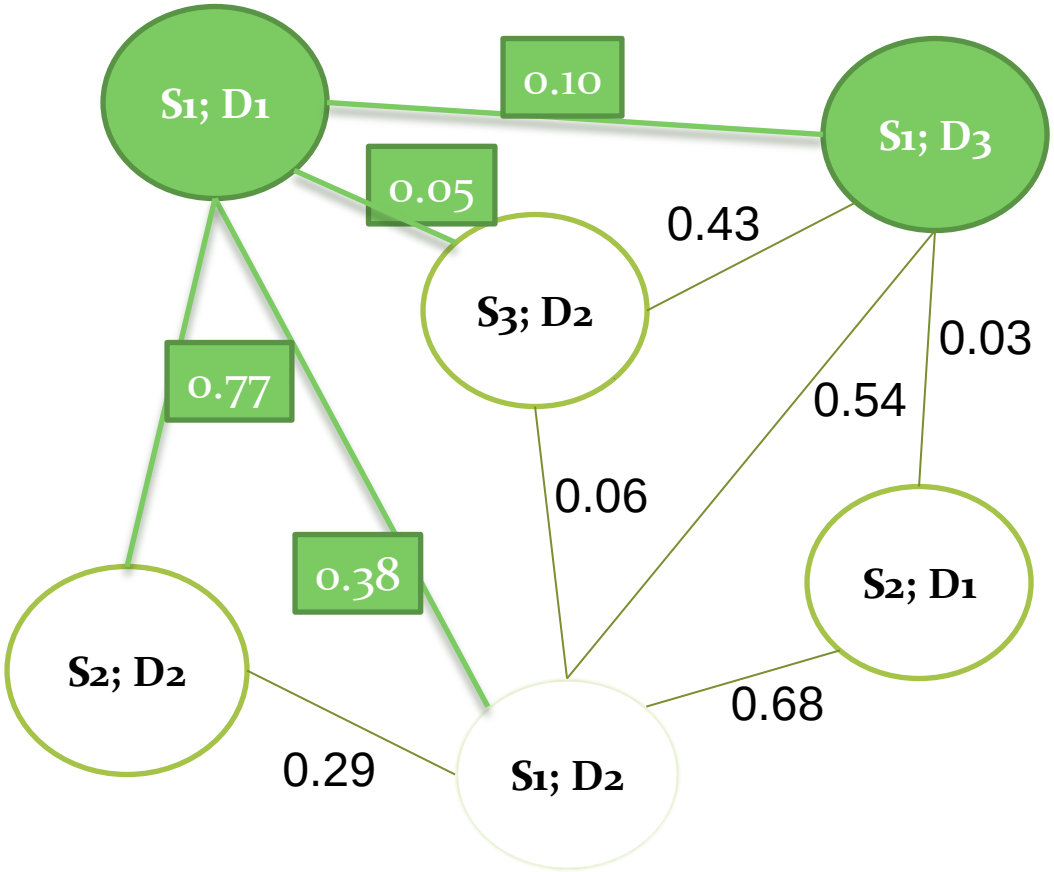


Sentence rank
S1; D2
–
–
–
–
–

Step by step

3.1. Graph and complex network measures

- Degree, avg. shortest path, clustering coefficient

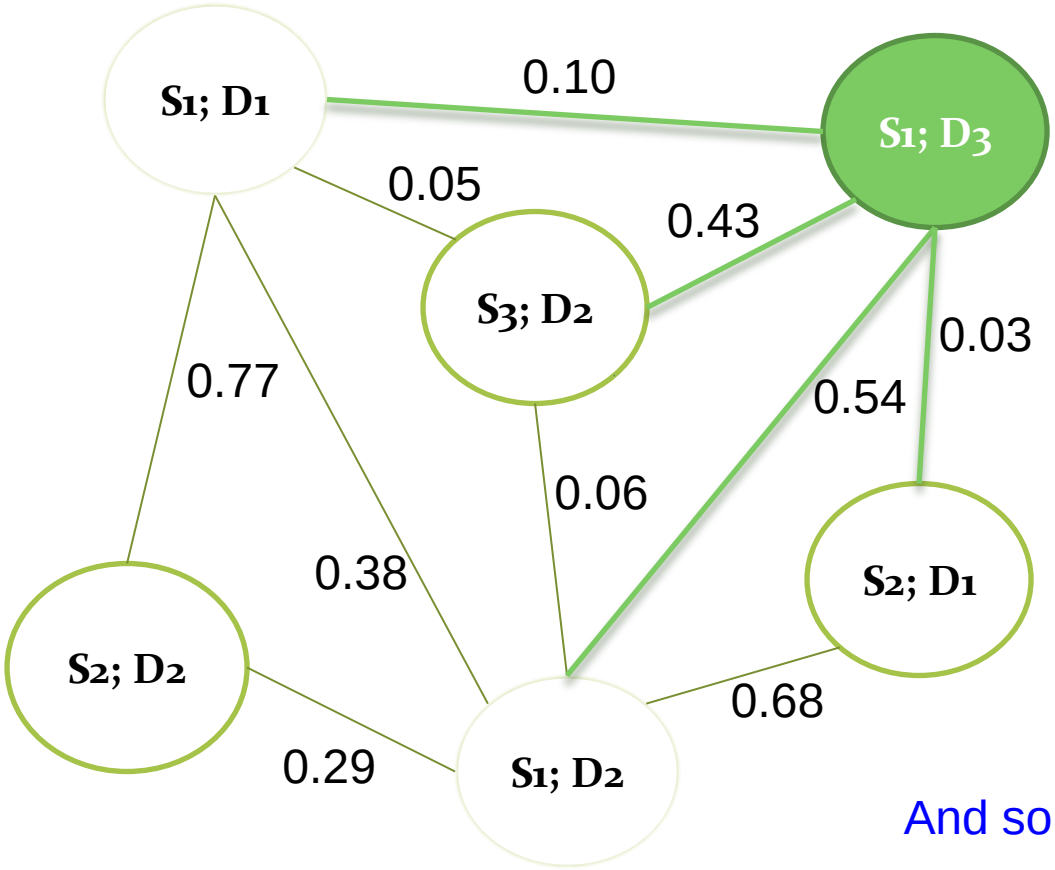


Sentence rank
S1; D2
S1; D1
–
–
–
–

Step by step

3.1. Graph and complex network measures

- Degree, avg. shortest path, clustering coefficient



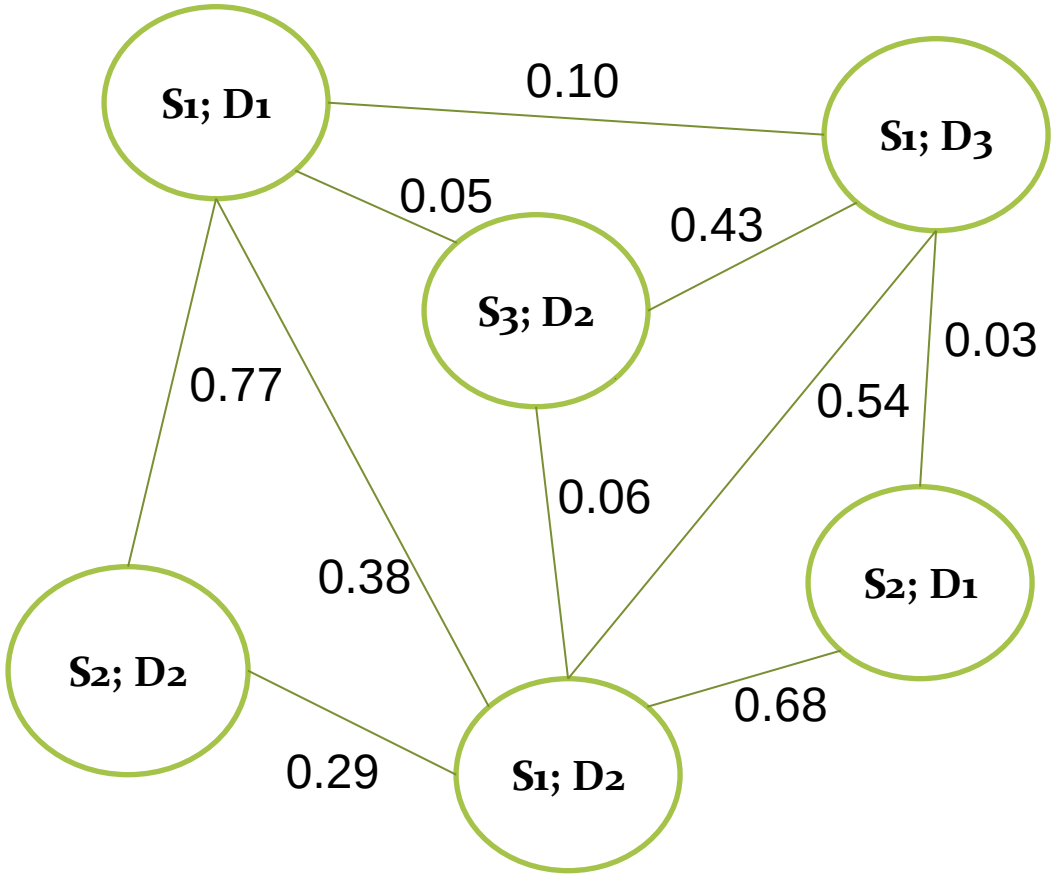
Sentence rank
S1; D2
S1; D1
S1; D3
–
–
–

And so on...

Step by step

3.2. Relationship maps

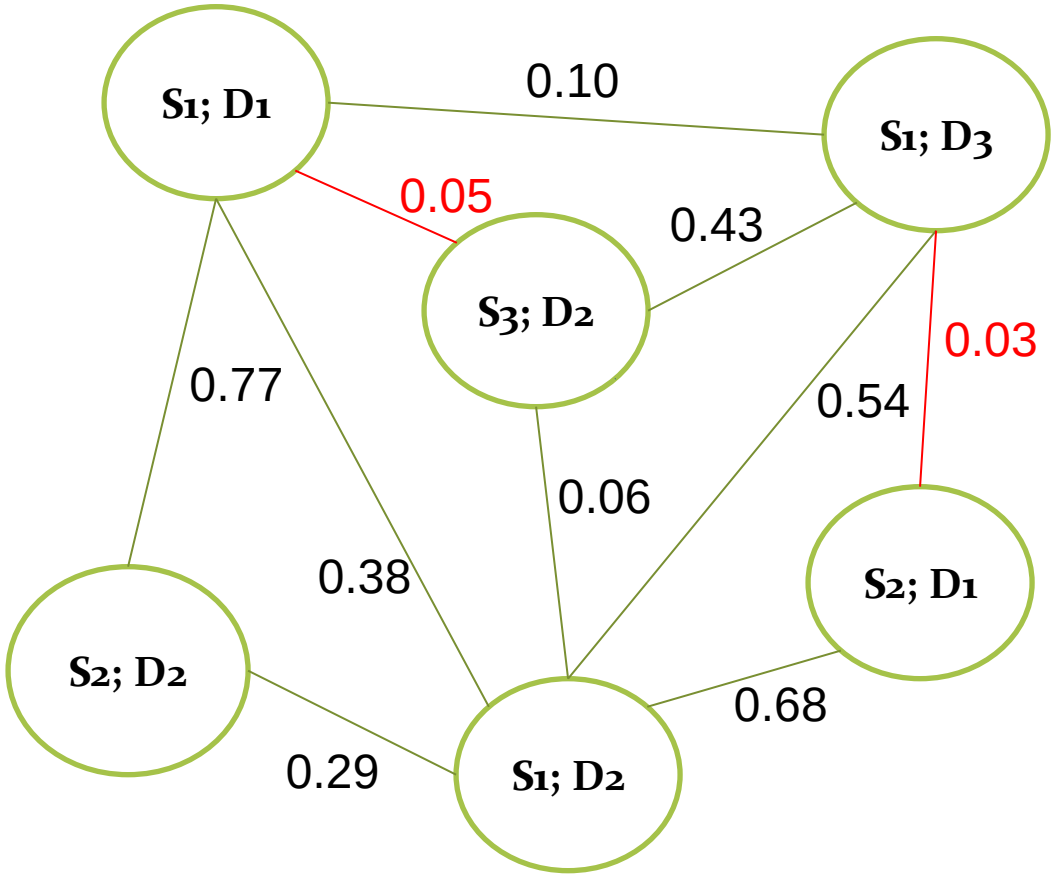
- Bushy path, depth-first path



Step by step

3.2. Relationship maps

- Bushy path, depth-first path

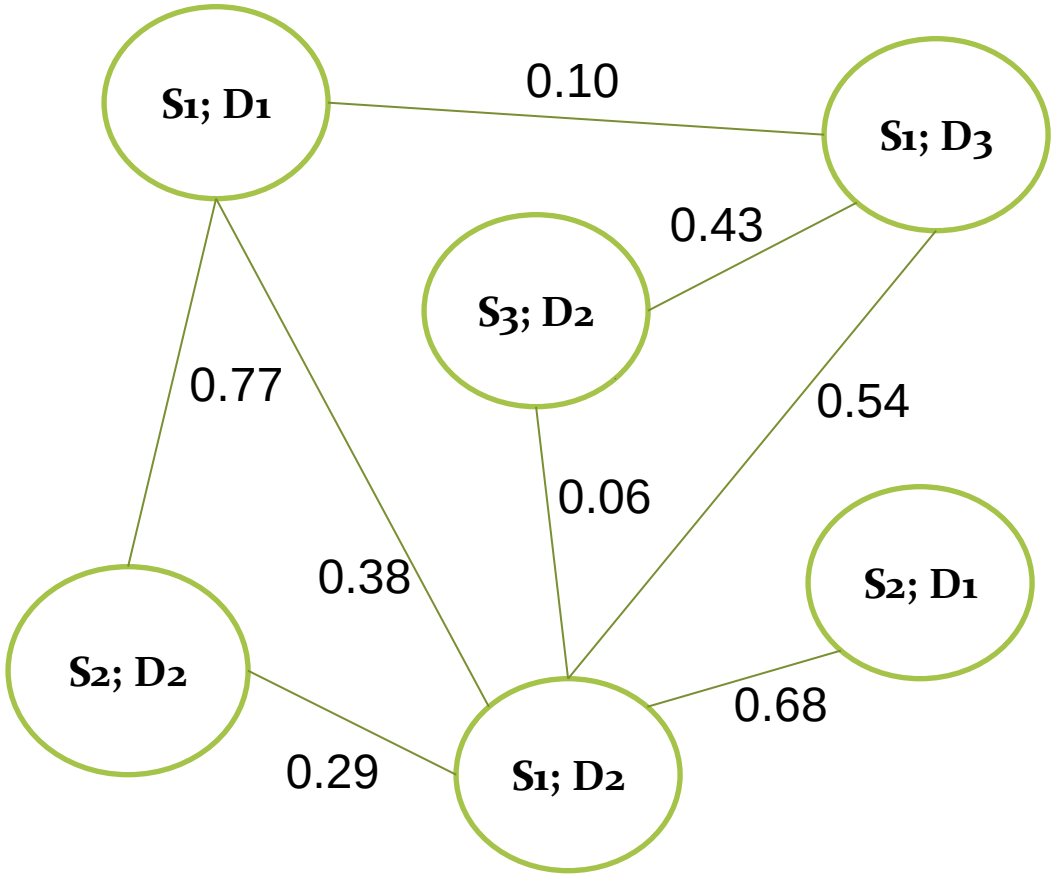


Map density parameter
*Keeping only the 1,5 * N best edges*

Step by step

3.2. Relationship maps

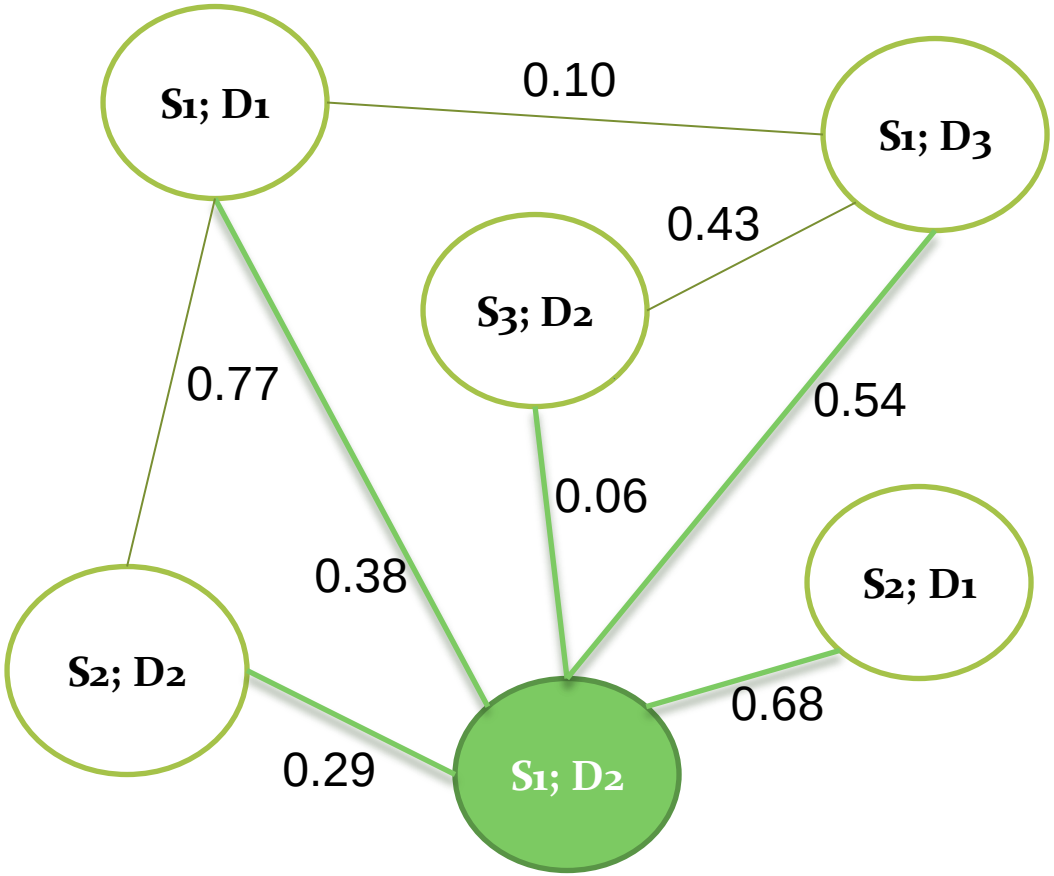
- Bushy path, depth-first path



Step by step

3.2. Relationship maps

- Bushy path, depth-first path

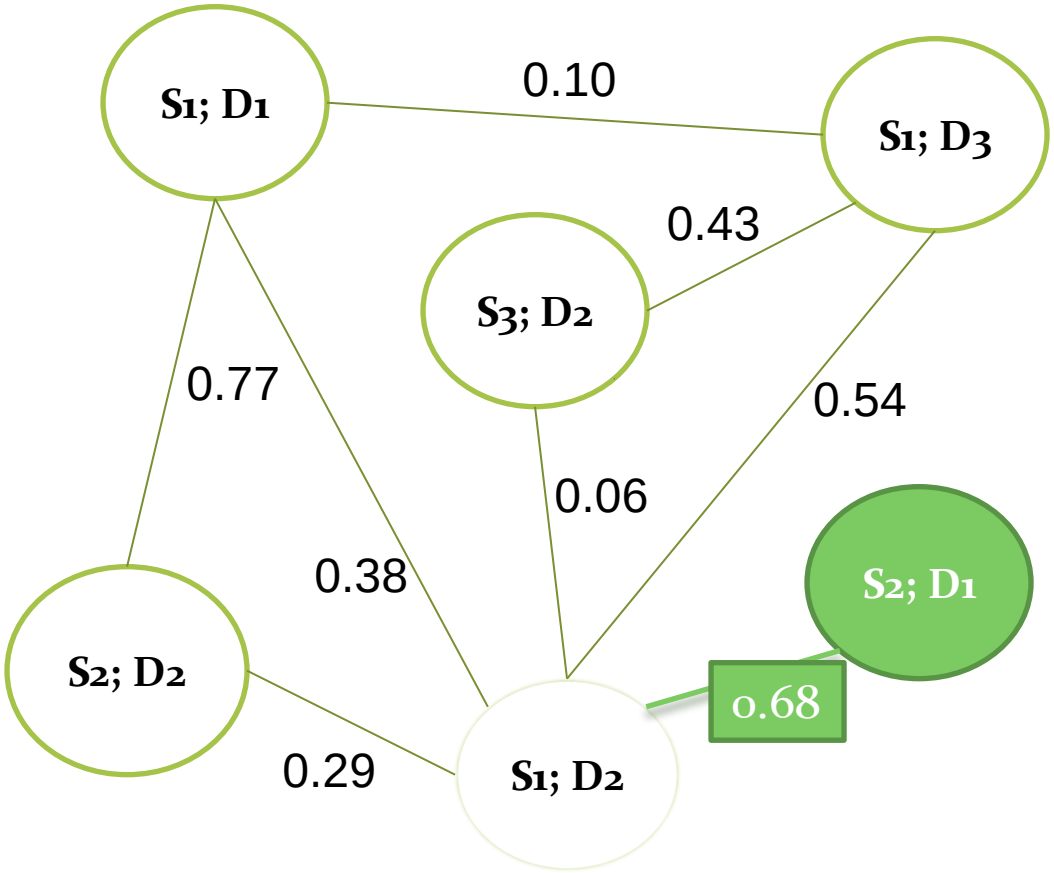


Sentence rank
S1; D2
–
–
–
–
–

Step by step

3.2. Relationship maps

- Bushy path, depth-first path

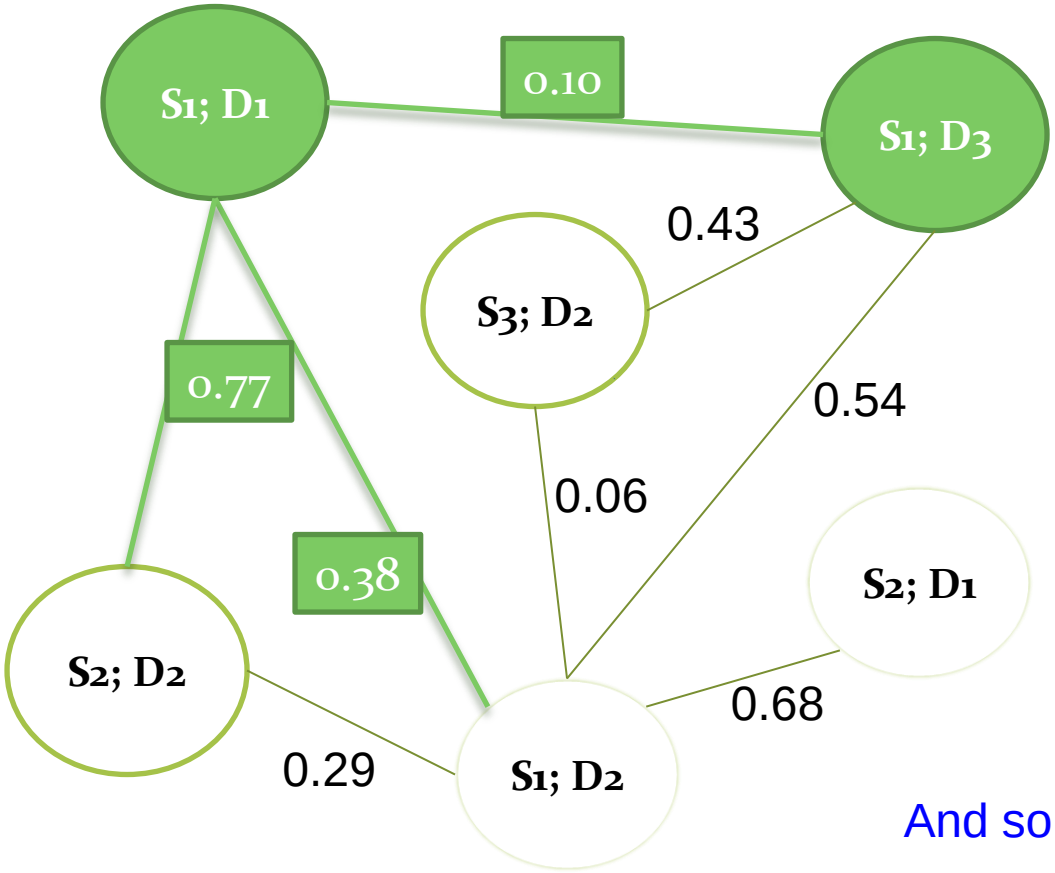


Sentence rank
S1; D2
S2; D1
–
–
–
–

Step by step

3.2. Relationship maps

- Bushy path, depth-first path



And so on...

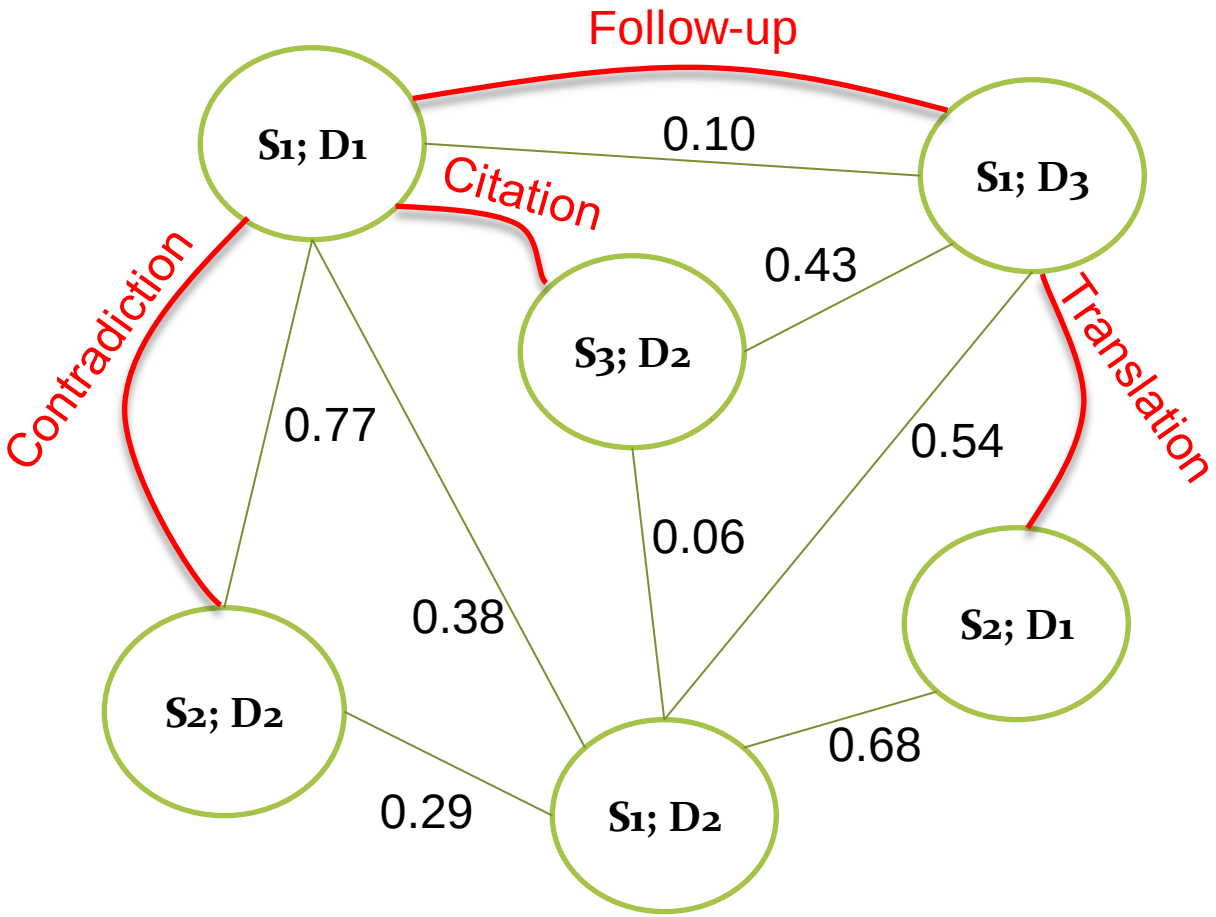
Sentence rank
S1; D2
S2; D1
S1; D1
–
–
–

Step by step

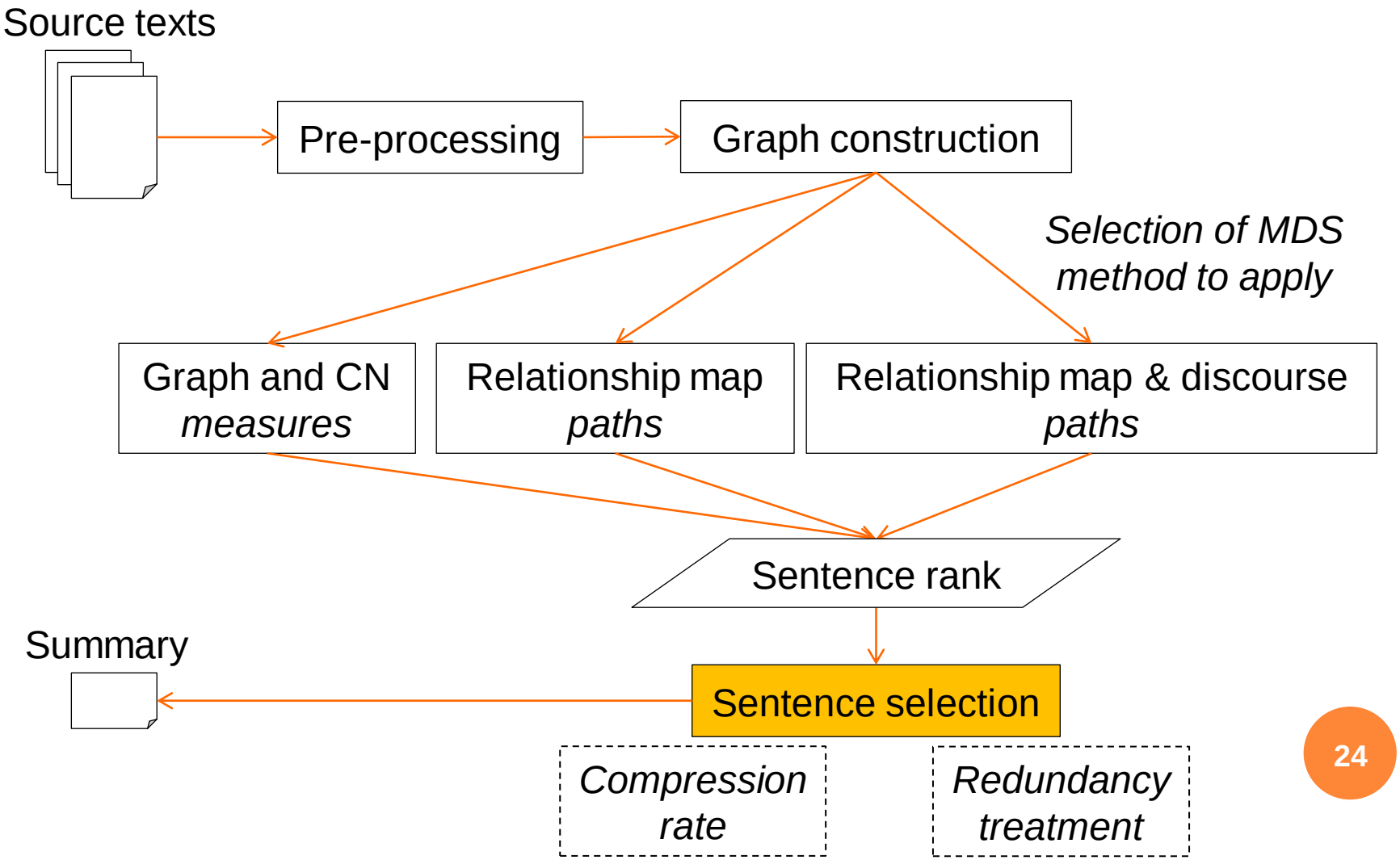
3.3. Relationship maps & discourse

- Bushy path, depth-first path

Adding CST relations



METHOD – OVERVIEW



Step by step

4. Sentence selection

- Starting from the best ranked sentences
  - Observing compression rate
  - Verifying the redundancy level in relation to previously selected sentences (using lexical similarity)
    - Redundant sentences are pruned

Methods

- Scientific foundations and expectations
  - Degree, shortest path, clustering coefficient, and bushy path: information centrality
  - Depth-first path: information centrality & information contiguity
    - Preference for redundant sentences (before pruning)
  - Discourse: meaning for more fine-grained decisions

Evaluation

- CSTNews corpus (Cardoso et al., 2011)
  - 50 clusters of news texts
  - Manual multi-document summaries
  - Manual CST annotation, with good agreement values

- Informativeness
  - ROUGE (Lin and Hovy, 2003)
    - Precision, recall and f-measure

- Comparison to other systems for Portuguese and to MEAD (Radev et al., 2001)

Results

- Degree is below CSTNews, still the best system
  - Statistically significance

System/Method	Precision	Recall	F-measure
CSTSumm	0.5547	0.5492	**0.5467**
Degree	0.5328	0.5037	**0.5155**
Shortest Path	0.5306	0.5009	0.5131
Bushy Path	0.4844	0.5397	0.5083
Bushy Path with CST	0.4844	0.5397	0.5083
Depth-first Path	0.4811	0.5340	0.5040
Depth-first Path with CST	0.4811	0.5340	0.5040
MEAD	0.5242	0.4602	0.4869
GistSumm	0.3599	0.6643	0.4599
Clustering coefficient	0.4671	0.4476	0.4560

Results

- Discourse only reinforces the graph-based results, not altering the results
  - As Louis at al. (2010) also claim

System/Method	Precision	Recall	F-measure
CSTSumm	0.5547	0.5492	0.5467
Degree	0.5328	0.5037	0.5155
Shortest Path	0.5306	0.5009	0.5131
Bushy Path	0.4844	0.5397	**0.5083**
Bushy Path with CST	0.4844	0.5397	**0.5083**
Depth-first Path	0.4811	0.5340	**0.5040**
Depth-first Path with CST	0.4811	0.5340	**0.5040**
MEAD	0.5242	0.4602	0.4869
GistSumm	0.3599	0.6643	0.4599
Clustering coefficient	0.4671	0.4476	0.4560

Results

- The 2 paths perform similarly

System/Method	Precision	Recall	F-measure
CSTSumm	0.5547	0.5492	0.5467
Degree	0.5328	0.5037	0.5155
Shortest Path	0.5306	0.5009	0.5131
Bushy Path	0.4844	0.5397	**0.5083**
Bushy Path with CST	0.4844	0.5397	0.5083
Depth-first Path	0.4811	0.5340	**0.5040**
Depth-first Path with CST	0.4811	0.5340	0.5040
MEAD	0.5242	0.4602	0.4869
GistSumm	0.3599	0.6643	0.4599
Clustering coefficient	0.4671	0.4476	0.4560

Results

- We are still far from human extractive results
  - As Genest et al. (2009) also show

Humans perform 30% better!!!

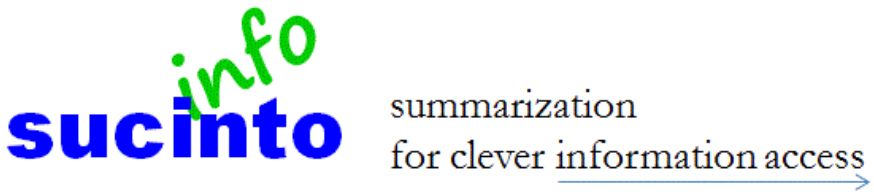
System/Method	Precision	Recall	F-measure
Humans	**0.6901**	**0.7216**	**0.7008**
CSTSumm	0.5547	0.5492	0.5467
Degree	0.5328	0.5037	0.5155
Shortest Path	0.5306	0.5009	0.5131
Bushy Path	0.4844	0.5397	0.5083
Bushy Path with CST	0.4844	0.5397	0.5083
Depth-first Path	0.4811	0.5340	0.5040
Depth-first Path with CST	0.4811	0.5340	0.5040
MEAD	0.5242	0.4602	0.4869
GistSumm	0.3599	0.6643	0.4599
Clustering coefficient	0.4671	0.4476	0.4560

Current and Future work

- Adaptation of one more Relationship Map method
  - Segmented bushy path: requires topic segmentation
- Human evaluation
  - Coherence and cohesion: manual evaluation
- Incorporation of other information processing tasks
  - Sentence ordering (Lima and Pardo, 2011)
  - Sentence simplification (Gasperin et al., 2010)
  - Sentence fusion (Seno and Nunes, 2009)

Graph-based Methods for Multi-document Summarization

- www.nilc.icmc.usp.br



Demonstration today!

A good automatic summary for 2 texts in the CSTNews corpus

> *A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.*

And a not so good one (for 3 texts in the corpus)

> *A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto. O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.*