European Centre for Soft Computing

“BisoNet” Generation using textual data

Marc Segond, Christian Borgelt

“BisoNet” Generation using textual data

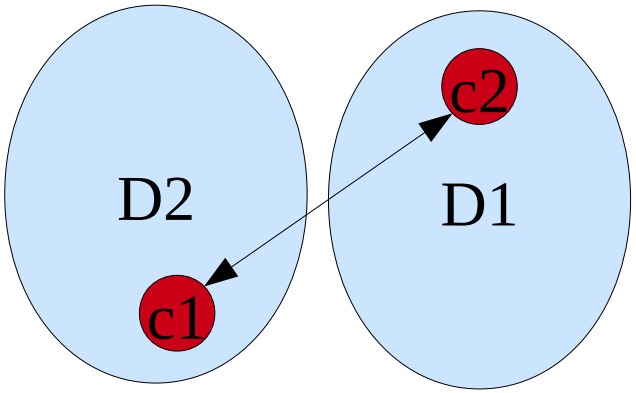
Content:

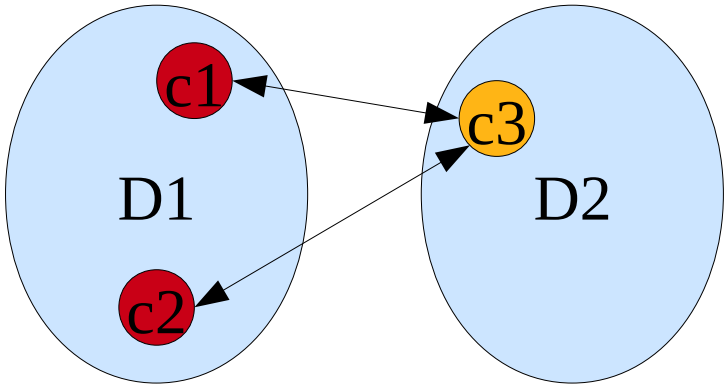
- Bisociations and BisoNets
- Data and pre-processing
- Creating nodes
- Creating and weighting links
- Benchmarks, results and further work

Bisociations and Bisonets

Definition of a Bisociation:

- A bisociation is a link L between two concepts c1 and c2 , which are unconnected given a specific context or view V . The concepts c1 and c2 may be unconnected, because they reside in different domains D1 and D2 (which are seen as unrelated in the view V), or because they reside in the same domain D1, in which they are unconnected, and their relation is revealed only through a bridging concept c3 residing in some other domain D2 (which is not considered in the view V).





Bisociations and Bisonets

BisoNet definition:

- Each concept (or, more generally, any named entity) gives rise to a node.
- Concepts that are associated (according to the classical paradigm of similarity or co-occurrence) are connected by an edge.
- Bisociations are then indirect connections (technically paths) between concepts, which cross the border between two domains (according to the previous definition).

Data and pre-processing

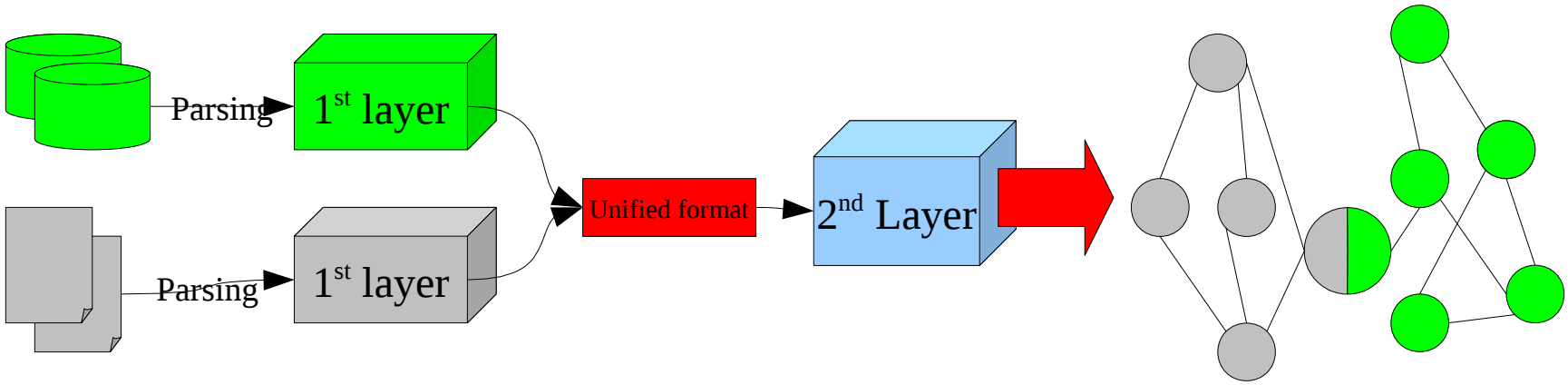
BisoNet generation requires:

- Components to access the original, usually heterogeneous, data sources.
- Methods for choosing the named entities that are to form the nodes of the BisoNet.
- Procedures for linking the nodes of a BisoNet and for endowing them with weights that indicate the association strength

Data and pre-processing

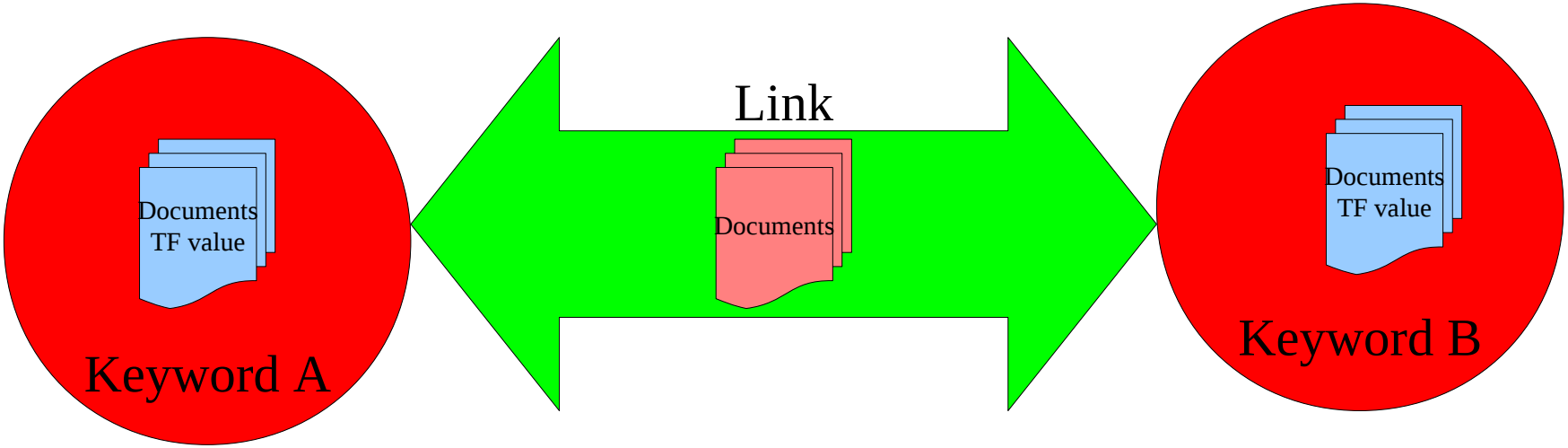
Our choices:

- A 2 layers framework:
  - First layer: customized parsers according to the databases to be read
  - Second layer: Bisonet generator
    - Generates a Bisonet from any set of parsers that gives data in the unified format in entry



Data and pre-processing

Actual structure of the Bisonet:



- Each node is associated to a keyword
- Each node contains a vector of links to relative documents
- Each node is weighted
- Each link contains a set of links to documents in which keywords A and B co-occur
- Each link is weighted

Creating nodes

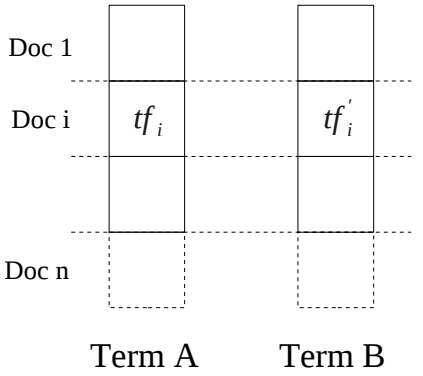
Nodes selection:

- Main terms are extracted using stemming and TFIDF
- TF calculation takes into account the difference between title, abstract and main text (when possible).
  - Words appearing in the title and abstract are weighted according to their occurrence frequency in the main text
- Nodes are weighted with their TFIDF value
- A node is a vector of the TF values of its keyword in each of the related documents.

Creating and weighting links

A link is created:

- Between two terms, according their co-occurrences in documents (if, at least, they co-occur in one document)
- A link between two terms is weighted with a similarity measure
  - 3 measures supported:
    - Cosine $\frac{A.B}{\|A\|\|B\|}$
    - Tanimoto $\frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}$
    - Bison



$$S(A,B) = \sum_i \sqrt[k]{tf_i \times tf_i'} \times \sum_i \left(1 - \frac{\left|\arctan(tf_i) - \arctan(tf_i')\right|}{\arctan(1)}\right) \qquad tf_i \in [0;1]; tf_i' \in [0;1]$$

Creating and weighting links

Specificity of the Bison measure

- Takes into account that two similar high TF values should give a highest similarity value than 2 low TF similar values
- Takes into account that it might be better to have two very similar but average TF values than one high and one low (e.g. two times 0.45 might be better than 0.3 and 0.7)

Benchmarks, results and further work

The Swanson benchmark:

- 8000 paper titles, taken from the PubMed database, published before 1987 and talking about either migraine or magnesium
- See if it is possible to re-discover relations between migraine and magnesium
- 2 “domains”: concerning magnesium and concerning migraine
- “Bridging concepts” will be keywords belonging to both domains

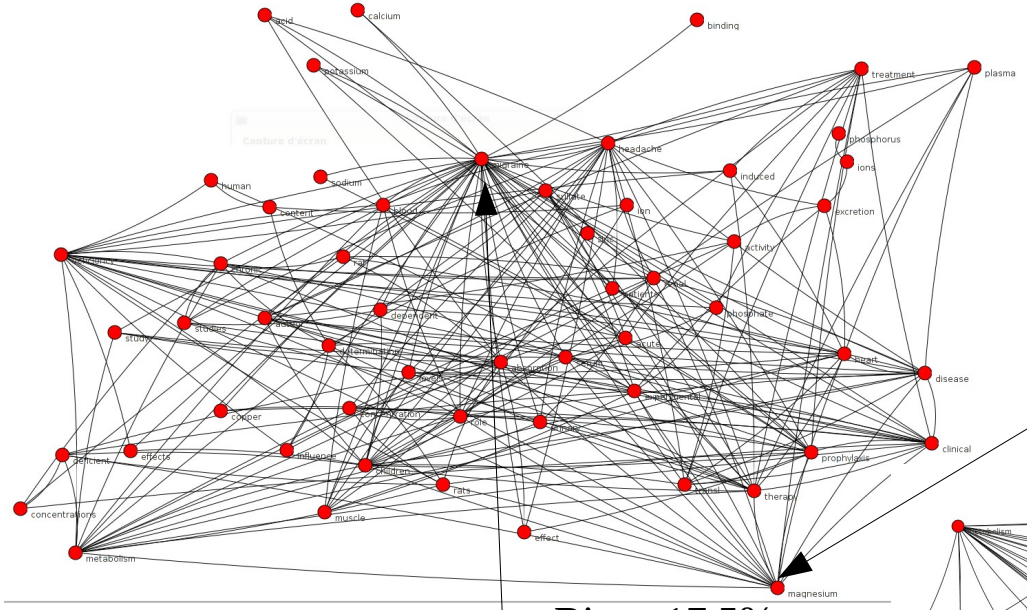
Benchmarks, results and further work

Comparison between similarity measures on the Swanson benchmark:

	% of (indirect) magnesium – migraine links kept		
% of highest ranked edges kept	Cosine	Tanimoto	Bison
5%	0%	0%	9%
17,5%	45%	47%	56%
100%	100%	100%	100%

European Centre for Soft Computing

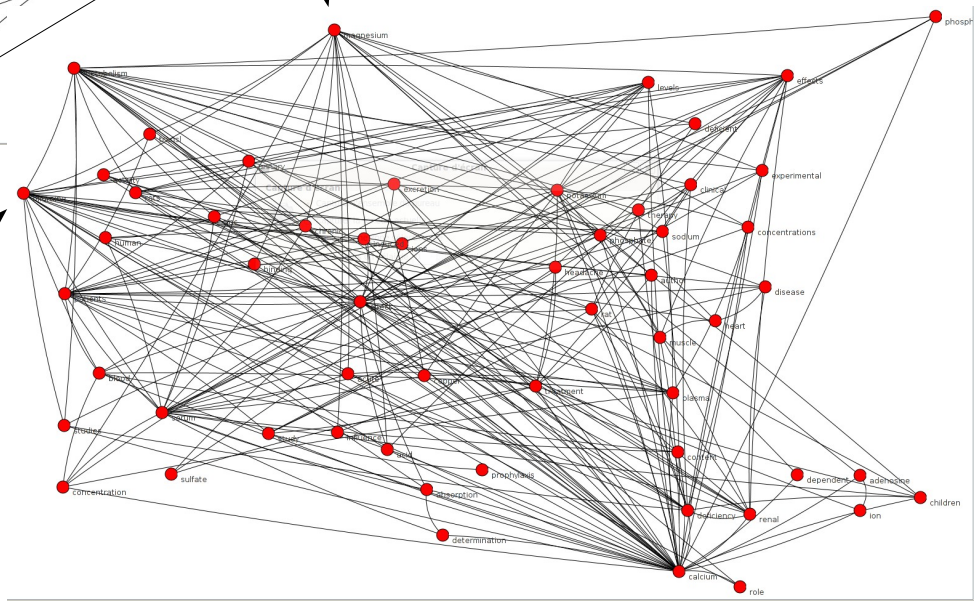
Benchmarks, results and further work



Bison 17,5%

Magnesium

Migraine

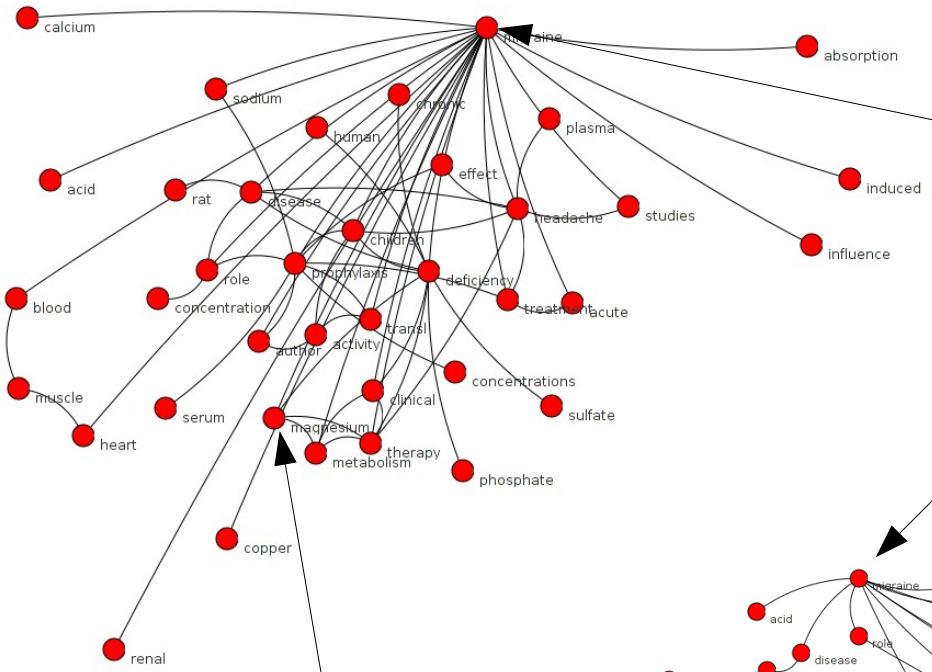


Tanimoto 17,5%

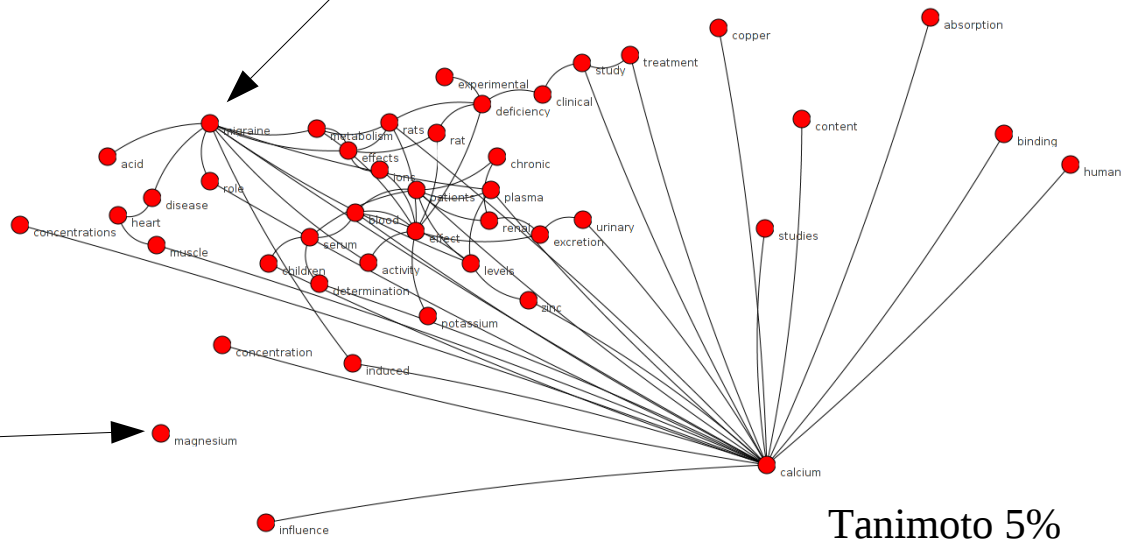
European Centre for Soft Computing

Benchmarks, results and further work

Bison 5%



Migraine



Magnesium

Tanimoto 5%

European Centre for Soft Computing

Benchmarks, results and further work

Results of the Swanson benchmark:

- We are able to easily re-discover the links between magnesium and migraine
- Cosine and Tanimoto measures give similar results, but the bison measure improves bisociations discovery
- The bison measure seems to highlight different links than the 2 other measures (e.g. calcium)
- Using this bison measure makes one able to use classical mining algorithms without tuning them to follow the “weakest paths”

Benchmarks, results and further work

An other benchmark: PubMed and FreeDB

- Two very different domains: biology and music:
- Data sources: PubMed and FreeDB
- Challenge: trying to merge two VERY different domains and see if there are any bisociations.
- Nodes are keywords extracted from titles and abstracts in PubMed and from titles and styles in FreeDB.
- Instead of “documents”, we talk here about “textual records”
- Same generation procedure as the Swanson benchmark

Benchmarks, results and further work

First results:

- There are bisociations! (but are they relevant?)
- Again, the Bison measure improves bisociations discovery

	% of cross-domain links kept		
% of highest ranked edges kept	Cosine	Tanimoto	Bison
5%	0%	0%	1,3%
25%	2,5%	3,9%	28%
50%	47,7%	49,6%	85,7%
100%	100%	100%	100%

Conclusion

What do we have here:

- A framework able to generate BisoNets from any textual data source
- Use of classical similarity measure as well as a custom measure adapted to bisociation discovery

What is still to be done:

- Testing with other data-sources
- More comparisons with other similarity measures
- Applying graph mining algorithms
- Real-life tests (really trying to solve actual problems, helping researchers to find unusual and interesting associations)