Thematic Cohesion: measuring terms discriminatory power toward themes

Clément de Groc* **Xavier Tannier†** **Claude de Loupy***

*Syllabs
26 rue Notre Dame de Nazareth, 75003 Paris, France
{cdegroc,loupy}@syllabs.com

†LIMSI-CNRS
Univ. Paris-Sud, 91403 Orsay, France
xtannier@limsi.fr

Abstract

We present a new measure of thematic cohesion. This measure associates each term with a weight representing its discriminatory power toward a theme, this theme being itself expressed by a list of terms (a thematic lexicon). This thematic cohesion criterion can be used in many applications, such as query expansion, computer-assisted translation, or iterative construction of domain-specific lexicons and corpora. The measure is computed in two steps. First, a set of documents related to the terms is gathered from the Web by querying a Web search engine. Then, we produce an oriented co-occurrence graph, where vertices are the terms and edges represent the fact that two terms co-occur in a document. This graph can be interpreted as a recommendation graph, where two terms occurring in a same document means that they recommend each other. This leads to using a random walk algorithm that assigns a global importance value to each vertex of the graph. After observing the impact of various parameters on those importance values, we evaluate their correlation with retrieval effectiveness.

Keywords: Thematic Lexicon, Random Walk, Graphs

1. Introduction

Lexicons and terminologies are key resources for natural language processing. They are useful in a wide variety of tasks such as text clustering or opinion mining. In this article, we focus on domain-specific resources, as the sample thematic lexicon shown in Table 1.

Building manually such resources is a long and expensive task, and semi-automatic processes using Web corpora or automatic translation tools can help reducing time and costs (Baroni and Bernardini, 2004; Kilgarriff and Grefenstette, 2003; Wang and Cohen, 2007). However, a manual validation step is still required when high quality resources are needed.

In this article, we present a new measure of *thematic cohesion*, to be applied to an existing thematic lexicon. This measure associates each term with a weight representing its discriminatory power toward a theme, this theme being itself expressed by a list of terms (a thematic lexicon). The measure is computed in two steps. First, a set of documents related to the terms is gathered from the Web by querying a Web search engine. We then produce an oriented co-occurrence graph, where vertices are the terms and edges represent the fact that two terms co-occur in a document. This graph can be interpreted as a recommendation graph, where two terms occurring in a same document means that they recommend each other. This leads to using a random walk algorithm (Page et al., 1999) that assigns a global importance value to each vertex of the graph.

This thematic cohesion criterion can be used in many applications. In bootstrapped, iterative construction of domain-specific lexicons, it can drastically reduce the time devoted to manual validation or help reduce topic drift through iterations. In computer-assisted translation, a cohesion value can provide the translator with hints concerning out-of-domain translations (de Groc et al., 2012).

For the sake of clarity, we define the problem of computing the thematic cohesion of a lexicon as follows: given a thematic lexicon $\mathcal{L}_T$ containing N terms, $\mathcal{L}_T = (t_1, t_2, \ldots, t_N)$, we aim at providing a weight vector $\mathbf{w}_{\mathcal{L}_T} = (w_1, w_2, \ldots, w_N)$ where each value w_i represents the discriminatory power of term t_i with respect to the theme T.

2. Gathering exogeneous knowledge

We use a standard Web search engine and collect, for each term t_i, a corpus C_i corresponding to the first M results from the search engine for the query "t_i".

We consider and compare two different units of information in our experiments: the web page and the snippet. With the entire web page, we take advantage of larger and richer contextual information, but downloading the documents is time-consuming and requires to clean the HTML files properly. On the other hand, snippets can be a good candidate for reducing noise and keeping a better focus on the specific domain.

Table 1: Extract of a thematic lexicon on astronomy.

afterglow	celestial coordinates	asteroids
dwarf stars	bow shocks	quasars
x rays	films	Einstein shift
red dwarf stars	auroral jets	space plasmas
accretion disks	solar atmosphere	Hubble telescope

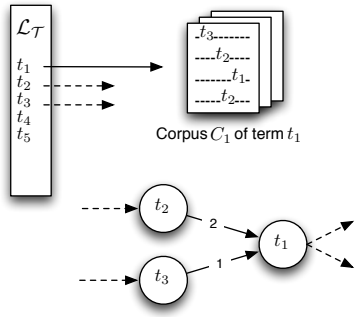


Figure 1: Example of subgraph built from the lexicon $\mathcal{L}_T$.

3. Thematic Cohesion Value

Considering a set of documents C_i retrieved by the term t_i, our assumption is that the importance of term t_i for the theme T depends on how many documents from the set C_i deal with this theme. For example, if the theme is astronomy, the term "*mercury*" is ambiguous (can be a planet, a Roman god, a chemical element, a singer), while "*telescope*" is not, and should get a high score for this theme.
However we do not have information concerning the theme of retrieved web pages. Rather than using a classifier for each theme, we will estimate the theme of documents by the number of terms from the initial lexicon that they contain. As a first definition, the weight w_i of a term t_i is the number of terms in the lexicon (t_i excluded) cooccurring with t_i in the corpus C_i:

$$w_i = \sum_{t_j \in \mathcal{L}_T^\dagger} n_{t_j,C_i} \quad (1)$$

where n_{t_j,C_i} is the number of occurrences of the term t_j in the corpus C_i, and $\mathcal{L}_T^\dagger = \mathcal{L}_T \setminus \{t_i\}$, the set of all terms of the lexicon, t_i excepted.
We normalize this score to obtain a value between 0 and 1:

$$w_i = \frac{\sum_{t_j \in \mathcal{L}_T^\dagger} n_{t_j,C_i}}{\sum_{k \in \{1,\ldots,|\mathcal{L}_T|\}} \sum_{t_j \in \mathcal{L}_T^\dagger} n_{t_j,C_k}} \quad (2)$$

Another way to see this formula is to represent cooccurrences of terms by an oriented graph $G =< V, E >$, where each vertice of V is a term of $\mathcal{L}_T$ and E is the set of edges, as illustrated by Figure 1. Each edge $e(t_i, t_j) \in E$ represents the occurrence of term t_i in C_j, weighted by its number of occurrences. Equation 1 then corresponds to the indegree of a term (Newman, 2004), while Equation 2 is its normalized version.

Note that this cohesion value differs from the existing TextRank algorithm (Mihalcea and Tarau, 2004). While Mihalcea and Tarau model term co-occurrences in a fixed-size window with a non-oriented graph, in our work, we consider that a term t_i *votes* for t_j when t_i appears in C_j, which is not a symmetrical relation, since C_i and C_j are distinct corpora.

Equation 2 still considers that all terms in the lexicon have the same influence on the weight of each term t_i. We now want to integrate the weights of terms t_j when computing the weight of t_i. The weight w_i of a term t_i increases with the number and weight of the terms from the lexicon that co-occur with t_i in the corpus C_i:

$$w_i = \sum_{t_j \in \mathcal{L}_T^\dagger} \frac{n_{t_j,C_i} \cdot w_j}{\sum_{k \in \{1,\ldots,|\mathcal{L}_T|\}} n_{t_j,C_k}} \quad (3)$$

Equation 3 is a random walk algorithm similar to PageRank (Page et al., 1999), and can be solved by the power iteration method. This method converges to a unique solution under two conditions (Langville and Meyer, 2005; Farahat et al., 2006): the graph must be strongly connected and the adjacency matrix of the graph must be stochastic (sums of all lines must equal 1). This is ensured by a normalization and a uniform teleportation vector (Page et al., 1999), leading to the following, final equation:

$$w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \sum_{t_j \in \mathcal{L}_T^\dagger} \frac{n_{t_j,C_i} \cdot w_j}{\sum_{k \in \{1,\ldots,|\mathcal{L}_T|\}} n_{t_j,C_k}} \quad (4)$$

where N is the number of vertices in the graph (*i.e.*, the number of terms in the lexicon) and α is a damping factor, traditionally set to 0.85.

Figure 2 shows a "toy" graph built with a few names of fruits. Edge direction is clockwise (Fekete et al., 2003); font size and intensity of colors represent the cohesion value.

Algorithm 1 sums up the computation of this thematic cohesion value.

4. Evaluation

4.1. Behavior

We first evaluate the influence of the number of documents fetched (M) and the initial lexicon size (N) on the output values, as well as the benefits of using snippets rather than the entire web page. For these experiments, we use four reference lexicons:

- Astronomy (2940 terms, *The Astronomy Thesaurus*[1])
- Statistics (2752 terms, *The ISI Glossary*[2])
- Medical-1 (2000 terms, *MeSH*[3])

[1] `http://msowww.anu.edu.au/library/thesaurus/`

[2] `http://isi.cbs.nl/glossary/`

[3] We use a subset of the MeSH (`http://www.nlm.nih.gov/mesh/`) from the UMLS 2008aa, and only consider Main Heading terms.

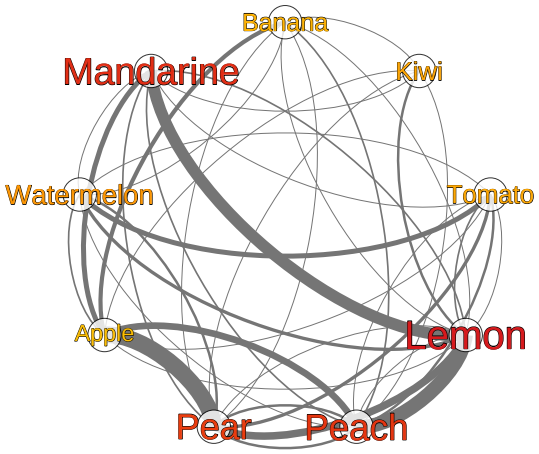


Figure 2: Case study: the "fruits". Edge direction is clockwise. Font size and intensity of colors represent the cohesion value.

Algorithm 1 Computing thematic cohesion value of a term for a theme

Input: $\mathcal{L}_T$: terms $t_i, i \in [1, N]$
M: number of documents to download for each query
α: damping factor

// Downloading the corpus
for all term $t_i \in \mathcal{L}_T$ **do**
  Query t_i to a search engine
  Download M documents/snippets as the corpus C_i
end for

// Initialization
for all term $t_i \in \mathcal{L}_T$ **do**
  $w_{i,1} = 1/N$
end for

// Computing the weights iteratively
$n = 1$
while (non-convergence) **do**
  for all term $t_i \in \mathcal{L}_T$ **do**
    $w_{i,n+1} = \frac{(1-\alpha)}{N} + \alpha \cdot \sum_{t_j \in \mathcal{L}_T^{\dagger}} \frac{n_{t_j,C_i} \cdot w_j}{\sum_{k \in \{1,\ldots,|\mathcal{L}_T|\}} n_{t_j,C_k}}$
  end for
  Weight normalization: $\sum_i w_{i,n+1} = 1$
  $n = n + 1$
end while
return w_n

- Medical-2 (2000 terms, *MeSH*[3])

A number of preprocessing steps have been applied to the lexicons to improve their quality for our evaluation. For the Astronomy and Statistics lexicons, we have cleaned up the terms by removing tokens between brackets or square brackets. While for the Medical lexicons, we simply discarded terms containing brackets or commas, such as *3-pyridinecarboxylic acid, 1,4-dihydro-2,6-dimethyl-5-nitro-4-(2-(trifluoromethyl)phenyl)-, methyl ester)*.

We used the Blekko[4] search engine, that provides a search API in english allowing 500 results per query, including URLs and snippets. An average snippet is made of 36 tokens from the retrieved web page, containing the term(s) of the query. No document at all were found for 14% of terms from Statistics lexicon and for less than 3% of terms from the other lexicons.

4.1.1. Number of documents

We first study the behavior of the thematic cohesion values while varying the number of documents/snippets fetched by query. We want to estimate the best trade-off between robustness (stable weights) and efficiency (time spent to download and parse documents). To this aim, we apply the algorithm described above with different values of M: between 10 and 100 for pages and between 50 and 500 for snippets. We then compare the weights obtained with the highest number of documents (*e.g.* 100 pages) with weights obtained with other numbers (*e.g.* 10, 20, 50 pages). We then compute Kullback-Leibler divergence KL (Kullback and Leibler, 1951) to measure the difference between two sets of weights. Kullback-Leibler divergence is a measure of the difference between two probability distributions p and q:

$$D(p,q) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)}$$

This leads to a positive score, and to zero for two identical distributions p and q.

Figure 3 shows the divergence with the entire web pages (3.a) and with the snippets (3.b). We can see that snippets offer a lower divergence and therefore a greater stability than web pages.

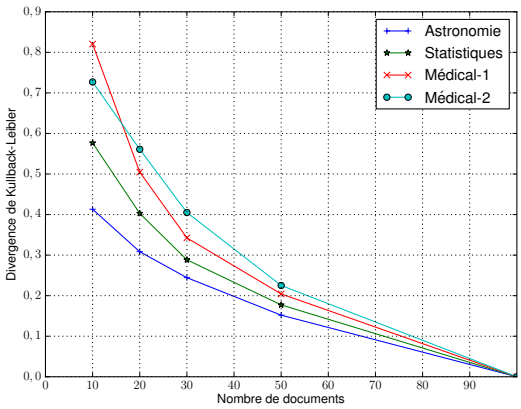
4.1.2. Web pages *vs* snippets

We now compare the usefulness of documents versus snippets to compute cohesion values. We estimated (with KL-divergence) the difference between results obtained with Web pages and those obtained with snippets, for all combinations of document numbers. Results are presented in Figure 4 for the "Astronomy" lexicon (others led to similar conclusions). Logically, the divergence decreases as the number of documents or snippets increases. Also, the divergence is quite stable for snippet numbers higher than 100 (below 0.2 for 100 snippets, 1.6 for 200 snippets). Only few snippets are enough to get close to the result obtained with 100 documents.
As snippets are much easier to obtain and to process, and lead to similar results than Web pages, we will use 200 snippets as input for the remaining of our work.

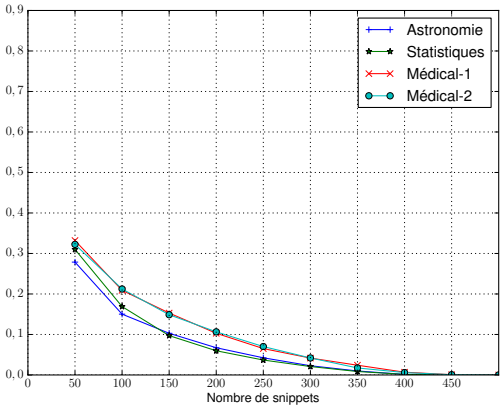
4.1.3. Size of the lexicon

The cohesion measure is essentially based on the appearance of the terms of the lexicon. Therefore, the size of the lexicon definitely influences the accuracy of the cohesion

[4] http://www.blekko.com



(a) Documents



(b) *Snippets*

Figure 3: Kullback-Leibler divergence between results obtained with the highest number of pages/snippets and a lower number.

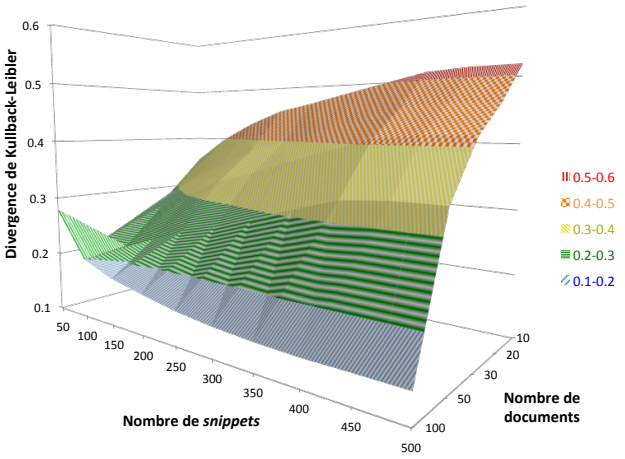


Figure 4: Kullback-Leibler divergence between results obtained with pages or snippets, for the lexicon "Astronomy".

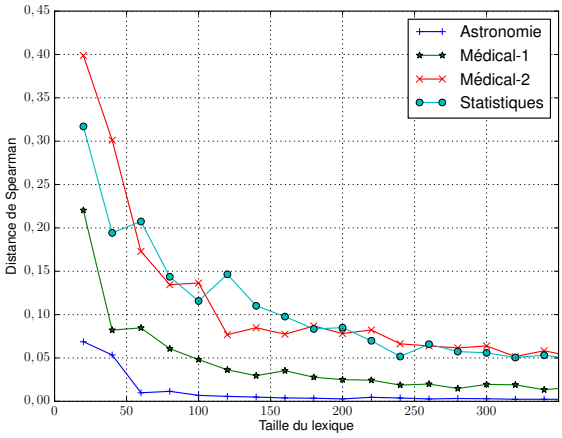


Figure 5: Normalized Spearman distance between term weights obtained for different sizes of lexicons.

values. Hence, we would like to evaluate the impact of the size of the lexicon on cohesion values. However, care must be taken as cohesion values cannot be compared directly using the KL-divergence as they are normalized by the number of terms in the lexicon (the sum of all weights makes 1). Therefore, we resort to compare the ranked lists obtained by sorting the lexicon terms by their thematic cohesion values. We performed several random selections of terms (from 20 to 1000 terms) while making sure that larger sets of terms included the smaller ones, and computed the weights of these lists. We then compared the evolution of term ranks in these different configurations. A configuration is stable if modifying the number of terms does not modify too much their global order. This is measured by the normalized Spearman's Footrule Distance (Diaconis, 1988; Dwork et al., 2001). Given two weighted lists A and B, with α and β the ranks of the terms in each lexicon, the normalized Spearman distance is:

$$D(\alpha, \beta) = \frac{\sum_{t_i \in \mathcal{L}_T} |\alpha(t_i) - \beta(t_i)|}{|\mathcal{L}_T|^2/2}$$

where $\mathcal{L}_T$ is our thematic lexicon.
Figure 5 shows the results of this experiment. Variation is important until 100 terms, but decreases quickly until 250, where the ranks become stable.

4.2. Relevance

Our last evaluation experiment aims at showing that our cohesion values correctly identify terms leading to relevant documents, and can hence be used to collect thematic corpora with higher precision. As a reference, we use the OpenDirectory (DMOZ)[5], a manually-fed web site directory where each page falls into one or several categories. To perform our evaluation, we need a search engine over the OpenDirectory and a set of thematic lexicons. We build both from the OpenDirectory by indexing documents from 340 categories of the second level of the DMOZ thesaurus

[5] `http://www.dmoz.org`

Table 2: Spearman correlation coefficient between prediction (thematic cohesion or *tf.idf*) and query result (average precision) for second-level categories of the OpenDirectory.

Measure	**Spearman ρ**	**Significance ($p < 0,05$)**
tf.idf	0.200 ($\pm$0.164)	32%
Thematic cohesion	0.434 ($\pm$0.208)	74%

in a search engine, and automatically extracting thematic lexicons of 200 terms from categories descriptions (see our companion paper (de Groc and Tannier, 2014)).

We compute cohesion values for each thematic lexicon using the Blekko Web search engine. Then, we issue each term as a query to our OpenDirectory search engine and compute the average precision of the set of retrieved documents. Average precision (AP) evaluates not only the relevance of retrieved documents, but also the quality of the ranking:

$$\text{AP} = \frac{\sum_{k=1}^{n}(\text{Precision}(R_k) \times \text{Relevant}(k))}{\text{number of relevant documents}}$$

where R_k is the set of retrieved documents, ranked from 1 to k, and Relevant(k) is a boolean function of the relevance of the document retrieved at rank k.

We then estimate Spearman's correlation coefficient between the average precision and weights produced by the thematic cohesion or by a baseline *tf.idf*. When close to -1 or 1, a Spearman coefficient show high anticorrelation or correlation, while 0 means that there is no correlation at all.

Table 2 presents Spearman's correlation averaged over all 340 categories. We also show the percentage of categories with a significant correlation value (p-value < 0.05).

We observe that *tf.idf* values are positively but weakly correlated with average precision, and that the correlation is only significant for 32% of all categories. On the other hand, the thematic cohesion offers a moderate to high correlation with high significance. This last result shows that our graph-based algorithm correctly identifies discriminative terms leading to relevant documents toward the theme.

5. Conclusion

We have presented a novel thematic cohesion measure that weights thematic lexicon terms according to their discriminatory power toward the theme. We believe that this measure can be useful to a wide panel of tasks and we plan to apply it to Web-based lexicon and corpora bootstrapping. Regarding the measure itself, we also plan to extend it to not only favour disciminative terms leading to relevant documents but terms leading to a large number of relevant documents.

6. References

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the LREC 2004 conference*, pages 1313–1316.

C. de Groc and X. Tannier. 2014. Evaluating Web-as-corpus Topical Document Retrieval with an Index of the OpenDirectory. In *Proceedings of 9th International Language Resources and Evaluation Conference (LREC)*.

C. de Groc, X. Tannier, and C. de Loupy. 2012. Un critère de cohésion thématique fondé sur un graphe de cooccurrences. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 183–195. ATALA/AFCP.

P. Diaconis. 1988. Group representations in probability and statistics. *Lecture Notes-Monograph Series*.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM.

A. Farahat, T. LoFaro, J.C. Miller, G. Rae, and L.A. Ward. 2006. Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201.

J.D. Fekete, D. Wang, N. Dang, A. Aris, and C. Plaisant. 2003. Overlaying graph links on treemaps. In *IEEE Symposium on Information Visualization Conference Compendium (demonstration)*, volume 5.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

S. Kullback and R.A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

A.N. Langville and C.D. Meyer. 2005. A survey of eigenvector methods for web information retrieval. *SIAM review*, pages 135–161.

R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 2004, pages 404–411.

M.E.J. Newman. 2004. Analysis of weighted networks. *Physical Review E*, 70(5):056131.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab.

R.C. Wang and W.W. Cohen. 2007. Language-independent set expansion of named entities using the web. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 342–350.