EPIQUE: Extracting Meaningful Science Evolution Patterns from Large Document Archives

Ke Li
LIP6, CNRS, Sorbonne Université
Paris, France
ke.li@lip6.fr

Hubert Naacke
LIP6, CNRS, Sorbonne Université
Paris, France
hubert.naacke@lip6.fr

Bernd Amann
LIP6, CNRS, Sorbonne Université
Paris, France
bernd.amann@lip6.fr

ABSTRACT

There is an increasing demand from domain experts for tools that assist them to extract information about the scientific progress and technological innovations from bibliographic archives such as the Web of Science, arXiv, PubMed, etc. Topic evolution graphs track the evolution of science by identifying and analyzing science evolution patterns like the emergence and decay of research topics or the split of one research topic into several subtopics, etc. Building such topic evolution networks for extracting meaningful evolution patterns is still a difficult task requiring the tuning of several technical parameters. In our demonstration, we present our prototype implementation of a generic topic evolution model for representing and filtering evolution patterns extracted from very large document archives.

1 INTRODUCTION

Revealing meaningful evolution patterns from document archives has many applications and can be used to synthetize narratives from datasets accross multiple domains, including new stories, research papers, legal cases and works of literature [12]. The study of science evolution can help *philosophers and historians* of science [10] to test their theories with data, *researchers* to position their work in its scientific context, *policy makers* to foster innovation and get key indicators for decision-making processes, *industry* to evaluate the potential for innovation and technological transfer, *librarians* to classify scientific documents, etc.

Scientific evolution can broadly be studied by adopting a cognitive view or a social view on evolution dynamics. The cognitive view emphasizes the shared knowledge and the change of ideas (Kuhn's approach [10]), whereas the social view takes account of authorship and social interaction (e.g., citation graphs) [7, 13]. Bibliographic archives often include both kinds of information and there also exist methods which also combine both views to study science evolution [8]. In the interdisciplinary EPIQUE project[1] we adopt the cognitive view for modeling science evolution and assume that the evolution only depends on the content of the documents. Whereas this choice clearly reduces the expressivity of our evolution model it also decreases the "social" bias and detects more easily possible interactions between scientific ideas and contributions independently of any particular scientific community.

The goal of topic evolution networks is to track complex temporal changes by epoch-wise topic discovery and temporal similarity graphs aligning topics of different epochs. Existing evolution network based frameworks mainly can be distinguished by the chosen topic extraction and alignment methods. [4] comes up with a method to enable a bottom-up reconstruction of the dynamics of scientific fields. They generate topics by word co-occurrence graphs and align inter-temporal topics by Jaccard similarity [9]. [1] generates topics by a Hierarchical Dirichlet Process (HDP) [14] and uses Bhattacharyya similarity [2], representing the gradual speciation and convergence similar to biologic evolution, for identifying topic alignments. The alignment process also applies (asymmetric) Kullback-Leibler divergence (KLD) for detecting topic split and merge. [11] introduces a novel approach to the early detection of research topics by using the Computer Science Ontology[2] to model research topics in the Rexplore system. They apply a Clique Percolation Method (ACPM) for analyzing the dynamics between existent topics. Other examples of science evolution studies explore how "cognitive science" as a field has changed over the last three decades [6] or analyze topic evolution patterns (split, merge and knowledge transfer) in the field of Information retrieval (IR) [5].

The goal of our work is to develop a general framework which is easier to use by domain experts who can ignore the details of the underlying topic analysis methods. The contributions presented in our demonstration can be summarized as follows:

- We propose a generic topic evolution model enabling the specification and extraction of meaningful topic evolution patterns independently of a particular topic extraction method.
- We define high-level measures for estimating the quality of the topic extraction process and for characterizing the structural and quantitative evolution of topics during a time period. This enables the experts to tune the topic extraction process and explore large topic evolution graphs by defining complex topic evolution patterns.
- We implemented a scalable prototype on top of Apache Spark for processing large scientific corpora containing millions of documents and finding meaningful topic evolution graphs for both stable topics and highly evolving ones.

2 TOPIC EVOLUTION MODEL

Topic evolution graphs: We consider a *corpus* C of time-stamped documents, a set of *periods* P and a set of terms V (*vocabulary*). Let $\mathcal{M} : 2^C \rightarrow 2^{\mathbb{R}^{|V|}}$ be a *topic extraction method* generating for a subset of documents $C' \subseteq C$ a set of (sparse) weighted term vectors $\mathcal{M}(C') \subseteq \mathbb{R}^{|V|}$. We denote by $C_p \subseteq C$ the corpus of documents with timestamp $p \in P$ and by $T_p = \mathcal{M}(C_p)$ the *topic descriptions* extracted from the documents C_p of period p using topic extraction method $\mathcal{M}$. A *topic* $t \in T_p$ is then defined by a couple $t = (d, p)$ where $d \in \mathcal{M}(C_p)$. We will denote by $t.d$ the *topic description* and by $t.p$ the *topic period*. Observe that topics from different periods may share the same description. For example in Figure 1, P has 3 periods: $p1$="$2000 - 2002$", $p2$="$2002 - 2004$" $p3$="$2004 - 2006$", T_{p1} contains topics 54 to 92, and topic $92 = (d, p1)$, where d is a weighted vector containing

[1] This work was funded by French ANR-16-CE38-0002-01 project EPIQUE.

[2] http://cso.kmi.open.ac.uk/

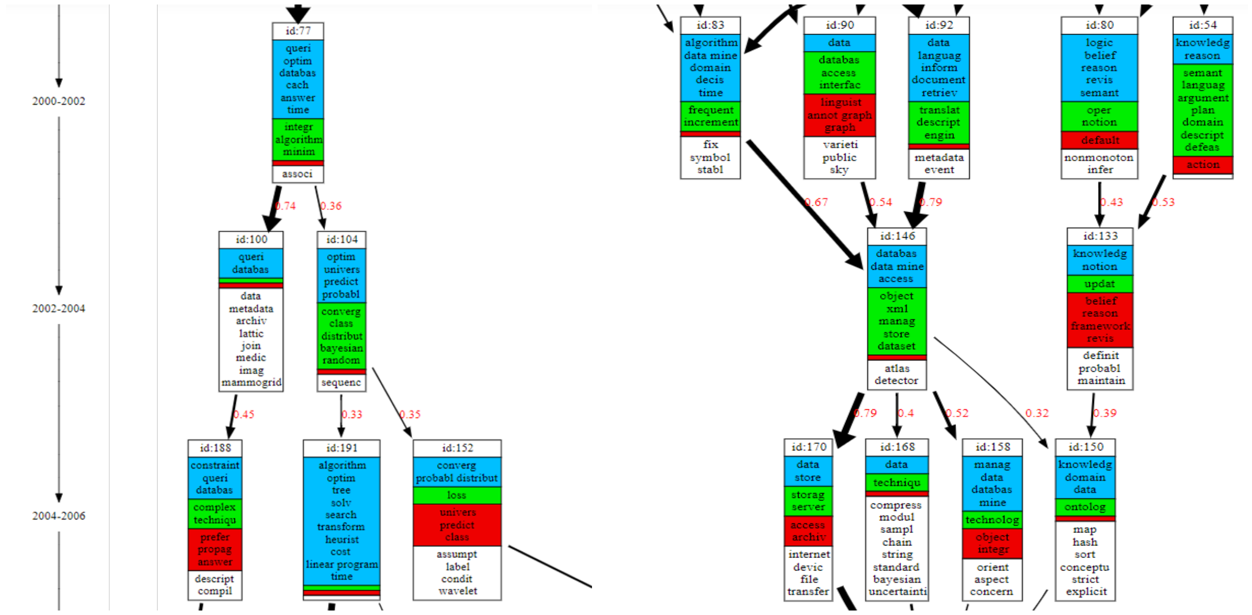


Figure 1: Pivot topics containing term "database" extracted from arXiv, green = emerging terms, blue = stable terms, red = decaying terms

terms "*queri*", "*optim*", "*databas*"... We define a *topic evolution function sim* : $T \times T \longrightarrow [0, 1]$ estimating the *similarity* between topics in T. For example in Figure 1, the similarity measure depends on the topic description and estimates their semantic proximity using *cosine* similarity. The similarity between topic 77 and topic 100 is $sim(77, 100) = 0.74$.

Based on the topic evolution function, we define a *topic evolution graph* as a directed labeled *multistage* graph $G_\beta = (T, E, sim, \beta)$ over topics T where the edges E connect all topics from consecutive periods with similarity higher or equal to some threshold β: $E = \{(t_i, t_j) \in T | sim(t_i, t_j) \geq \beta \wedge t_j.p = t_i.p + 1\}$.

Topic labeling: For visualization, we assume that all topics t of some evolution graph G_β are labeled by the top-k highest weighted terms in the topic description $t.d$. Let $t.l$ be the top-k highest weighted terms in $t.d$ and $t.l_p \subseteq t.l$ and $t.l_f \subseteq t.l$ be the subsets of *past* and *future* terms which appear, respectively, in the ancestor topics and in the descendant topics of t. Then, the terms in some topic vector $t.l$ are partitioned into the following four subsets of :

- *emerging* future terms $t.l_e = t.l_f - t.l_p$ which do not exist in past topics,
- *decaying* past terms $t.l_d = t.l_p - t.l_f$ which do not exist in future topics,
- *stable* terms $t.l_g = t.l_p \cap t.l_f$ which exist in the past and the future topics of t, and
- *specific* terms $t.l_s = t.l - (t.l_p \cup t.l_f)$ which neither exist in the past nor in the future topics of t.

The quadruple $[t.l_e, t.l_d, t.l_g, t.l_s]$ is called the *term label* of t.

Figure 1 shows two snippets of a single topic evolution graph extracted from the arXiv[3] corpus for the category DB (databases). Although the number of documents of category DB are limited, the generated graphs still generate meaningful evolution patterns. Emerging terms are shown in green boxes, decaying term boxes are colored in red, stable terms which exist both in ancestor topics and in descendant topics are grouped in blue boxes and specific terms which appear only in current topic are in white boxes. The thickness of edges reflects the similarity between topics. Several topics in both subgraphs contain the term "database" and we can observe different evolution patterns. The left hand graph shows that in period 2002-2004, topic 77 ("databases, queries, optimization, integration") split into topics 100 and 188 ("databases, queries and constraints") and topics 104, 191, 152 ("prediction, probability, random"). The right subgraph covers the same period with topics related to "data mining" (83), "data access interfaces" (90), "information retrieval" (92), "logics, semantics" (80) and "knowledge, reasoning" (54). The first three topics converge in 2002-2004 into a single topic on "object, xml, store, data mining" (146) which splits in the period of 2004-2006 into "storage servers" (170), "data mining and management" (158) and "knowledge and ontologies" (150).

Pivot evolution graphs: Threshold β strongly influences the complexity of the obtained evolution graphs. It is easy to see that $G_{\beta'}$ is a subgraph of G_β for all $\beta' \geq \beta$ and G_0 is the complete graph connecting all topics of two consecutive periods. More exactly, higher β values generate more "linear" graphs with many isolated topics, whereas lower values generate more complex graphs containing a variety of potentially interesting structures. Observe also that, whereas the pivot graph complexity of the same topic increases with decreasing β, high β thresholds might still generate complex pivot graphs and vice-versa. Analyzing science evolution by using topic evolution graphs then becomes a complex task which consist in computing and visually exploring multiple graphs for different β values. To solve this problem, we propose a different approach which allows users to formulate *filtering queries* for selecting interesting *subgraphs* with meaningful measures from a set of evolution graphs defined by a set of β thresholds. For this, we decompose topic evolution graphs into the set of all connected subgraphs defined by all paths containing a given topic t (one graph per topic). More formally, a *pivot evolution graph* $G_\beta(t) = (T', E', sim, \beta)$ *of topic* t in G_β is the subgraph of G_β which contains t and all *ancestors* and *descendants* of t. The subgraph of $G_\beta(t)$ containing all nodes which are reachable

[3] https://arxiv.org/

from t by a path is called the *future* of t, denoted by $F_\beta(t)$, and the subgraph of nodes which can reach t through a path is called the *past* of t, denoted by $P_\beta(t)$. The couple (t, β) is called a *pivot topic* with pivot graph $G_\beta(t)$, future $F_\beta(t)$ and past $P_\beta(t)$. It is easy to see that if t_1 appears in the future (past) of t_2, then the future (past) of t_1 is a subgraph of the future (past) of t_2 and t_2 appears in the past (future) of t_1. This property can be exploited to filter topics *wrt.* future and past topics (see the definition of Connection Filters below).

The evolution of topics within their evolution graphs can be characterized by the following metrics:

- The *liveliness* $live(G_\beta(t))$, of a pivot topic (t, β) is defined by the diameter (longest path length) of its pivot graph $G_\beta(t)$. A high liveliness value describes a long living topic, whereas a value equal to 0 corresponds to an isolated topic without ancestors and descendants. The liveliness $live(P_\beta(t))$ of t in its past estimates the "age" of t *wrt.* the first period, whereas $live(F_\beta(t))$ returns the "life expectation" of t (in its future).
- The *relative evolution degree* $revol(G_\beta(t))$ of a pivot topic (t, β) is defined by the average topic dissimilarity (edge) weight in $G_\beta(t)$. A low relative evolution degree states that most topics are connected to very similar topics, *i.e.*, most topics in $G_\beta(t)$ evolve slowly. On the other hand, a high value signifies that most topics have an important "semantic gap". By definition, $revol(G_\beta(t)) \geq \beta$.
- The *pivot evolution degree* $pevol(G_\beta(t))$ of a pivot topic (t, β) is defined by the average dissimilarity of all topics in $G_\beta(t)$ with respect to the pivot topic t. A low pivot evolution degree signifies that the pivot topic does not evolve a lot (all other topics are similar), whereas a high value indicates that the pivot topic evolves rapidly .
- The *split degree* $split(G_\beta(t))$ of a pivot topic (t, β) is defined by the average outdegree of $G_\beta(t)$. A low value signifies that the topics evolve along linear paths and a high value signifies that the topics split into several future sub-topics.
- The *convergence degree* $conv(G_\beta(t))$ of a pivot topic (t, β) is defined by the average indegree of $G_\beta(t)$. A low value signifies that many topics depend on a single parent topic and a high value signifies that many topics are the result of the fusion of past topics.

Evolution Pattern Filters: The previous evolution metrics characterize the the evolution of a topic in some evolution graph G_β. Combined with other filters on the topic labels and the graph structure, it is possible to filter pivot topics satisfying rich evolution patterns within a set of evolution graphs G_{β_i}, $1 \leq i \leq n$.

Term Filters select pivot graphs with respect to the pivot topic labels. In particular, they can be applied to filter pivot graphs *wrt.* to their emerging, decaying, stable, and specific terms.

Temporal Filters allow experts to filter the pivot topics situated within a certain time period.

Pattern Filters can filter topics by their pivot graph structure along their liveliness, split degree and convergence degree.

Evolution Filters are applied to filter topics by their relative and pivot evolution degrees.

The previous filters are applied to sets of pivot topics and can be combined with the following other kinds of operators:

Connection Filters are binary operators which select all pivot topics that are connected to at least one pivot topic in some other set of topics.

Temporal Projection allows to restrict structural, evolution and connection filters to the past or the future of the pivot topics.

Set Operators allow to combine two sets of topics by union, intersection and difference.

Ordering Operators sort pivot topics by their attributes, such as the topic period, its liveliness, evolution degree etc.

3 WORKFLOW AND IMPLEMENTATION

Figure 2 illustrates the overall workflow which takes as input a corpus of documents split into several, possibly overlapping time periods (the same document might appear in two periods).

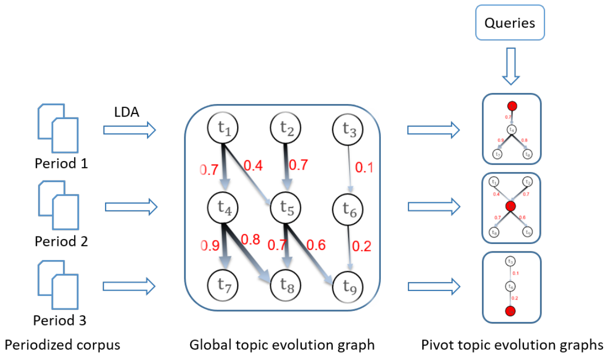


Figure 2: Topic evolution model of EPIQUE

All documents within a period are processed by LDA [3] to generate a set of topics which are aligned to produce a single topic evolution graph G_{β_0} for some small alignment threshold β_0. This global evolution graph is then transformed into n families of pivot evolution graphs defined by a set of alignment thresholds $\beta_i > \beta_0$, $1 \leq i \leq n$. Each family contains the pivot graphs $G_{\beta_i}(t)$ of all pivot topics (t, β_i). The final database contains $n \times |T|$ pivot graphs where $|T|$ is the number of topics in G_{β_0}. These graphs can then be queried using the filters defined in Section 2.

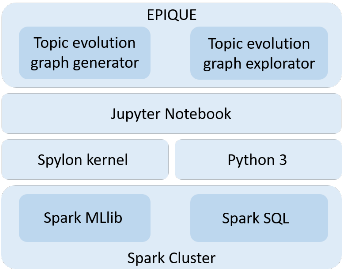


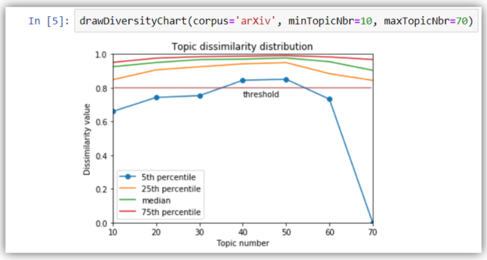
Figure 3: EPIQUE web application architecture

Figure 3 gives an overview of the architecture of our web application implemented on top of Apache Spark and Jupyter Notebook. The entire process to study science evolution over a corpus is split into two steps for building the pivot evolution graphs and for interactively exploring these graphs. Each step corresponds to a separate user interface. The evolution graph generation is implemented in Scala and exectued through the Spylon[4] kernel. Evolution graph exploration uses a standard Python kernel to take advantage of advanced Python 3 graphical user interface libraries for facilitating user interaction.

[4] https://github.com/Valassis-Digital-Media/spylon-kernel

4 DEMONSTRATION

Our EPIQUE prototype allows the audience to easily and intuitively generate high-quality evolution graphs and explore them. Among the corpora we have prepared in several domains, our demonstration focuses on the evolution of computer science based on the ArXiv corpus. We propose two interactive demonstration scenarios[5].



```
In [5]: drawDiversityChart(corpus='arXiv', minTopicNbr=10, maxTopicNbr=70)
```

Figure 4: Screenshot: topic diversity evaluation

Scenario 1 The audience selects or uploads a corpus of documents with a vocabulary of terms pre-processed by an on-line text-mining tool Gargantext[6] and specifies the time periods through sliding window over a global time period. Then, the LDA topic model is generated for each period. LDA requires a vocabulary and a number of topics to be generated. This number obviously influences the diversity of the resulting topics. Therefore, the application first generates a set of topic models for different topic numbers per period. The user can then visualize the diversity of the extracted topic models (topic dissimilairty distribution) and choose the model with the highest diversity for each period. A topic diversity distribution for different topic numbers is reported as shown in Figure 4 and, for example, by observing the 5th percentile values (blue line), the user can retain one of the two models (40 or 50 topics per period) that achieves 95% of pairwise dissimilarities above 0.8.

Then, the topics of consecutive periods are aligned and all pivot topic evolution graphs are generated along with their main temporal, structural and evolution indicators: *liveliness*, *split degree*, etc. All topic labels are also generated automatically in this step.

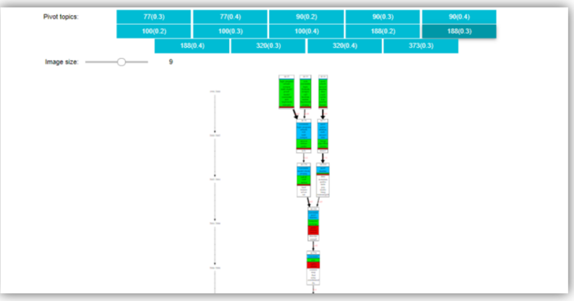


Figure 5: Screenshot: pivot topic evolution graph visualization

In the next step, the user specifies its exploration goal through an intuitive declarative query-by-example interface (as shown in the demonstration video[5]) and visualizes pivot topic evolution graphs as shown in Figure 5. These graphs are pre-computed in the last step to ensure fast query answer display.

We showcase a search for topic graphs containing a given term (*e.g.*, database) or set of terms suggested by the audience. Besides the topic content, the audience can search for topic graphs on their shape as well. We also prepared 10 predefined query templates for a typical shapes of high interest such as (i) topics that split in distinct 5+ years long branches, (ii) topic graphs with low *relative evolution degree* and high end-to-end *pivot evolution degree*. We also demonstrate more complex queries combining several query templates to build, for example, *concept drift* queries looking for pivot topics that contain emerging terms originating from other, "older" topics which are not part of their past pivot subgraph.

Scenario 2 In the second scenario, we will provide the audience with the possibility to prepare their own corpus using the Gargantext service which is also part of the EPIQUE project. Gargantext includes a number of bibliographic archives like Pubmed, Web of Science, etc. and allows to create domain specific document collections and vocabularies which are then processed by the same workflow as in Scenario 1.

5 FUTURE WORK

In the next step, we intend to optimize the computation of pivot topic evolution graphs and exploit the LDA document-topic matrix for enriching the analysis. Additionally, we plan to integrate other topic extraction methods than LDA. This prototype will also be used to validate our evolution model with philosophers of science to define and extract complex evolution patterns from different scientific domains.

REFERENCES

[1] Victor Andrei and Ognjen Arandjelović. 2016. Complex temporal topic evolution modelling using the Kullback-Leibler divergence and the Bhattacharyya distance. *EURASIP Journal on Bioinformatics and Systems Biology* 2016, 1 (2016), 16.

[2] A. Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35 (1943), 99–109.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] David Chavalarias and Jean-Philippe Philippe Cointet. 2013. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one* 8, 2 (2013), e54847.

[5] Baitong Chen, Satoshi Tsutsui, Ying Ding, and Feicheng Ma. 2017. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics* 11, 4 (2017), 1175–1189.

[6] Uriel Cohen Priva and Joseph L. Austerweil. 2015. Analyzing the history of Cognition using Topic Models. *Cognition* 135 (2015), 4–9.

[7] Eugene Garfield. 1955. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122, 3159 (1955), 108–111.

[8] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting Topic Evolution in Scientific Literature: How Can Citations Help?. In *ACM Conference on Information and Knowledge Management*. ACM, 957–966.

[9] Paul Jaccard. 1912. The Distribution of the Flora in the Alpine Zone.1. *New Phytologist* 11, 2 (1912), 37–50.

[10] Thomas S. Kuhn, Otto Neurath, and Thomas Samuel Kuhn. 1994. *The Structure of scientific revolutions* (2nd ed., enlarged ed.). Number ed.-in-chief: Otto Neurath ; Vol. 2 No. 2 in International encyclopedia of unified science Foundations of the unity of science. Chicago Univ. Press, Chicago, Ill.

[11] Angelo A. Salatino, Francesco Osborne, and Enrico Motta. 2018. AUGUR: Forecasting the Emergence of New Research Topics. In *ACM/IEEE on Joint Conference on Digital Libraries*. ACM, New York, NY, 303–312.

[12] Dafna Shahaf, Carlos Guestrin, Eric Horvitz, and Jure Leskovec. 2015. Information Cartography. *Commun. ACM* 58, 11 (2015), 62–73.

[13] Xiaoling Sun, Jasleen Kaur, Staša Milojević, Alessandro Flammini, and Filippo Menczer. 2013. Social Dynamics of Science. *Scientific Reports* 3 (2013), 1069.

[14] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*. 1385–1392.

[5] see http://www-bd.lip6.fr/wiki/site/recherche/projets/epique/demo/start for a video demonstration

[6] https://gargantext.org/