Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 39 (2016) 138 – 143

www.elsevier.com/locate/procedia

TFC 2015– TRIZ FUTURE 2015

Ideality & Bio-Inspired based collaborative bibliographic search method

Fayemi Pierre-Emmanuel[1,2], Duci Stefano[1,3], Fayolle Thomas[4,5], Nicolas Maranzana[2], Bersano Giacomo[1*]

[1] *Active Innovation Management, 37, rue des sources, Antony, France 92160*
[2] *Arts & metier Paristech, 150 boulevard de l'hopital, Paris, France 75013*
[3] *University of Bergamo, via Salvecchio 19, Bergamo, Italy 24129*
[4] *Ikos Consulting, 155, rue Anatole France, Levallois-Perret 92300*
[5] *École nationale supérieure d'informatique pour l'industrie et l'entreprise, 1 Rue de la Résistance, Évry, France 91000*

* Corresponding author. Tel.: +33608156809; *E-mail address:* g.bersano@aim-innovation.com

Abstract

Bibliographic search is a common entry point to any scientific work. It aims at gathering enough information to support the incoming thoughts. It is usually an unstructured process relying on both information retrieval and serendipity.

Considering the search as an unstructured process, it could easily become a time consuming never ending process.

In order to maximize the outcome of bibliographic search with limited time and resources, an approach is proposed. It focuses on how to structure such search without contradicting its need of serendipity thanks to different features. The first relies on the concept of Ideality. It aims at identifying and structuring the information to look for. The second feature focuses on identifying the scope and depth of the analysis related to any articles constituting the bibliography. These aspects are tackled thanks to an objective/subjective grading system with handling of the results inspired by the behavior of social insects. Another feature inspired by patent search (precision/recall indexes) focuses on identifying when to stop the article curation. Finally, parts of all these features can be coupled in order to generate a bibliographic search map. This map represents a means to share an overview of the work being done, allowing another person to take ownership of the bibliographic search. This feature provides the opportunity for structuring and strengthening collaborative bibliographic search approaches.

Peer-review under responsibility of Scientific committee of Triz Future Conference

Keywords: Bibliography; Collaborative; Ideality; Bioinspiration; Information Retrieval;

1. Introduction

From the beginning of any scientific work, it is common to invent every relevant document that will contribute to its process. This task, called preparing a bibliography, is used both to give a fair credit to the work of previous authors in the research field and to explicit the scientific positioning.

Bibliographies can be enumerative, systematic, descriptive, analytical, historical or textual [1]. Irrespectively of their type, both bibliographic searches [2] and collaborative bibliographic searches turn around three main activities: Achieving, writing and consulting a bibliography.

This article focuses on the first step of the process, the achievement of a bibliography. After introducing how it is usually done, the article provides an original model that structure the achievement process, allowing its outcome to be integrated within a collaborative framework. This model and the components on which it relies (ideality, swarm intelligence concept, precision and recall indexes and mapping) will then be described. Finally, two case studies will be presented, one from a scientific context, through an ontology's state of the art student project, the second from an industrial context.

2. Achieving a bibliography

As the bibliographic search has the purpose to establish and profit from an aggregated set of information, it is part of collaborative information retrieval (IR) [3]. TRIZ and information retrieval, with approaches such as [4, 5] found common ground on the concept of functional searches.

Peer-review under responsibility of Scientific committee of Triz Future Conference
doi:10.1016/j.procir.2016.01.179

2.1. Common process for achieving a bibliography

Looking for relevant scientific and technical information implies several sub-steps.

The first one is the transformation of the research question into a list of keywords and descriptors likely to concentrate a relevant content. This compiled thesaurus is the entry point to indexing and searching for documents. The thesaurus should be adjusted in order to reduce noise by incorporating words that are as uncommon and unambiguous as possible and by precisely defining questioning criteria. However, the thesaurus should also avoid silence, i.e., having too few results, by using synonymous or more generic terms.

The second step is database selection. It is usually made under the consideration of the typology of the databases (primary or secondary information), their scope, their operators (Boolean and/or linguistic), and their specific language.

2.2. Actual limitations

A bibliographic search is usually done by a single researcher, compiling documents according to his own judgment and selection criteria [6]. The common process it relies on, is the same that has been used since the age of Alexandra [7], an unstructured one, which allows the required serendipity to occur. The quality of a bibliography will thus be strongly correlated with the researcher's information retrieval skill.

If the process to perform a bibliographic search has not evolved since ancient times, our world has noticeably changed. Over time, our systems, whether scientific or industrial, have shown an increasing complexity [8]. That inherent complexity has led to the integration of more distant knowledge and to increase the interdisciplinary of research and/or development teams [9].

As bibliographic search is intrinsic to any scientific approach, it also has to be thought from a collaborative perspective. Existing methods that help researcher to perform bibliographic search are not meant to provide the needed degree of interactivity for researchers to collaborate on achieving it. Therefore, a limitation has been identified.

Gull [10] reports that when two teams were trying to agree on the relevance of retrieved documents, they agreed that 1390 documents were variously relevant to a set of 98 questions, but disagreed on a further 1577 documents, and the disagreements were never resolved.

In this paper, the problem of assessing a document's relevance is supported through the IFR ideal final result from TRIZ. Through this concept, it has been possible to define a metric for each document, based on objective information and subjective information.

3. Collaborative bibliographic search model

In order to provide a collaborative bibliographic search means, a new framework has to be developed. The model should offer:

- A structural frame that would not exclude the needed serendipity.
- A high degree of interactivity, for collaboration to be effective.
- Communication means for information exchange between researchers.

In recent years, several means focusing on allowing collaborative bibliographies have arisen. Nevertheless these tools, called reference management software or services, are innovative shared work-environment and do not provide any guidelines or procedural steps on the achieving process. Therefore, a collaborative bibliographic search model is proposed.

3.1. Ideality to define relevance

Most of the solutions that come from engineers are based on trade-off. Trade-off means partial fulfillment of opposed objective without excessive deterioration of all parameters of the system under study; it allows easier solution finding. According to Altshuller, every technical system strives towards an ideal state. Ideal Final Result, or IFR, is about picturing this ideal state by overcoming current technological limitations. Ideality is reached when an action is fulfilled without the need of a physical system [11]. That Ideality is characterized by the following formula:

$$I = \frac{\sum Fu}{\sum Fn + \sum Fc} \qquad (1)$$

With: I being the level of proximity to Ideality, Σ Fu being the sum of the useful functions, Σ Fn being the sum of the harmful functions and Σ Fc being the costs generated by the system.

In the context of bibliographic search, the concept of ideality can be used to initially and precisely define the information to look for. The purpose is to explicitly state all questions the bibliographic search will answer. By doing so, the process defines what is called the 'ideal paper'. Once the required answers outlined, they can be used as a reading grid, allowing the reader to screen documents in a more effective way.

The 'ideal publication' is a model. It represents the publication that, if written, would have entirely covered the intended bibliographic search. The ideal publication can also be seen as the review article the researcher may write at the end of a bibliographic search process. It is thus matching the Ideality formula by covering all the actual needs (Fu) while maintaining low costs (Fc), as the useful information is condensed in one document and without triggering any negative effect (Fn).

Every piece of information processed during the bibliographic search can then be analyzed through the ideal publication framework. The more a document provides valuable information, the more close to the ideal publication it is. According to this fact, a scoring system can be achieved. This Ideal Publication Index is composed of two sub-indexes:

- Subjective-Index (S-Index)
- Objective-Index (O-Index)

The Subjective-index, aims at picturing how targeted questions are answered by a single article. It represents the percentage of addressed questions, the quality of the disclosed information and the triggered scope aperture (has the article enlarged the research scope in any new relevant way?). Those three criteria are assessed on a 30 points score system (as shown in table 1).

Table 1. S-Index

Criteria	Score	Coefficient
Percentage of addressed questions	0-10	2
Quality of information	0-10	3
Scope aperture	0-10	1

The Objective-index, aims at assessing the article environment. The first criterion is the number of cumulative citations of the document. The more the document has been cited the more impact it has on the community it belongs to. The second criterion is the journal's impact factor or editors' renown; the higher it is the more efforts are required for researchers to publish their paper, meaning the publication quality is supposedly higher. Another criterion is the publication date. More recent publications provide two benefits:

- They are more likely to provide a better filtering of past references.
- They provide a better overview of the research question by considering/embedding previous state of the art in the research development process they disclose.

Regarding these two benefits, closeness of a given publication to the ideal publication seems to be correlated to the recentness of its publication. The last criterion is the type of publication, ranging from 0 to 10, according to the rigor of the peer review process.

Table 2. O-Index

Criteria	Score
Number of citations	(Citations/x(curent y-publication y))/x
Journal's Impact Factor	(IF/highest IF in the field)*10
Date	0-10
Type of publication	0-10

With X being the required variable for the number of citations criterion to reduce the score to ten.

The overall score obtained is a mark out of a hundred. Has the ideal publication would score a hundred, every publication can be seen as a percentage of this ideal publication.

The scoring system allows both a fast assessment of every curated publications during the former steps of the bibliographic search and an easy consultation from other contributors.

3.2. Stigmergy for the coordination

While they have various cognitive abilities, social insects are characterized by the fact that they show collective problem-solving capabilities that can only be explained by the contribution of their group behavior [12].

Since the beginning of their work in the 1950's, termites, ants, bees and wasps are the most emblematic social insect species.

Their organizational capability has led insect populations to develop features such as division of labor, specialization, collective regulation, plasticity, mass action responses or building of complex architectures [13].

These extraordinary capabilities are not – as it was previously thought [14] – based on a centralized information process, but are managed without any apparent means of regulation, by a self-organizational process in which agents are governed by a set of simple rules [15]. Evidence of the ecological success of social insects can be found almost everywhere [16], this special feature has been studied by both philosophers and entomologists. Among the generated concepts there is consensus that stigmergy [17] is ruling some of these self-organizational processes. This particular type of communication "workers are stimulated by the performance they have achieved."

Stigmergy is characterized by the fact that it is an indirect, non-symbolic form of communication mediated by the environment, coupled by the fact that stigmergic information remains on a local scale [18].

One example of a stigmergic communication based mechanism is the formation of recruitment trails in ant colonies. When scouting for food, ants follow a structured process. Scouts usually start looking randomly for food sources. When an ant finds a source, it returns to the nest, marking the ground with pheromones. This trail will lead other randomly scouting individuals to the source. As every individual will release pheromones on its way back to the nest, the track will be reinforced over time. This mechanism has led to more than 8 optimization algorithms and it is now known that TCP protocols which is mainly used for internet traffic works in the same way that ants are looking for food [19].

With the similar purpose of optimization in mind, Tero et al. have reconstructed the Tokyo railway's network thanks to expanding amoeba-like blob and oat flakes dots of size according to the daily transported population [20].

In the context of bibliographic search these approaches could lead to both an optimization of resources and an adaptive task allocation, thanks to their decentralized systems base. The purpose of using these concepts within the model is to allow information seeker to better investigate their curated documents. Thanks to the ideal publication concept, every document has a score. Following the stigmergy/slime concept, seekers should allocate time spent on every document, according to the score set. The higher the score is, the more carefully circum-investigation (retrieving other publications from the same author(s), from the same laboratory, etc.) should be done.

3.3. Patent search

Bibliographic search and patent search share two similar needs: the serendipity of their process and the fact that they can become a never ending process. It is thus interesting to see how patent search techniques can contribute to collaborative bibliographic searches.

The mechanism of screening documents through the user interaction can be related to information retrieval techniques with feedback [21]. There are several types of relevance feedback techniques, such as direct feedback, indirect and blind relevance feedback. The general idea of relevance feedback (RF) is to involve the user in the retrieval process, the user gives feedback on the relevance of documents in an initial set of results. If the results are a set of documents, the user marks them as relevant or irrelevant and the information retrieval system elaborates the selection proposing a new set of documents. In many cases, the information retrieval system uses Rocchio algorithms or similar to propose a new set of documents that is likely to have higher precision (the fraction of retrieved documents that are relevant).

Common relevance feedback techniques try to maintain the interaction of users as low as possible, limiting the relevance evaluation to a binary option (relevant and irrelevant). In this paper, the evaluation of relevance can be more detailed thanks to the collaborative aspects of the search. In fact, while a seeker rapidly screens the documents marking them as relevant or not relevant, the evaluator will further read them to associate a more detailed degree of relevance. This degree of relevance is seen by the users through the visualization tool. The more the degree of relevance of a document is, the more users will concentrate around it, looking into citations, papers.

3.4. Monitoring performances of the bibliographic search

The evaluation and monitoring of the bibliographic search can lead to its smart management. One problem of bibliographic searches is in fact related to the time needed to complete it.

Basically, we can consider that the end of the process should be identified as the moment where no more interesting documents are likely to be identified in an acceptable amount of time.

Monitoring performances of the information retrieval process can be done with available metrics already known in literature. In the simplest case, where relevance is considered as a binary concept, the retrieved documents can be classified as:

- true positives: retrieved documents that are relevant;
- false positives: retrieved documents that are not relevant;
- true negatives: not retrieved documents which were not relevant;
- false negatives: not retrieved documents which were relevant.

This classification is largely used to define metrics for the evaluation of an information retrieval system. In the case of unranked retrieval results, commonly used parameters are recall, precision and F-measure [21].

Precision is the fraction of retrieved documents that are relevant:

$$Precision = \frac{\#(true\ positives)}{\#(retrieved\ items)} \qquad (2)$$

Recall is the fraction of relevant documents that are retrieved:

$$Recall = \frac{\#(true\ positives)}{\#(relevant\ items)} \qquad (3)$$

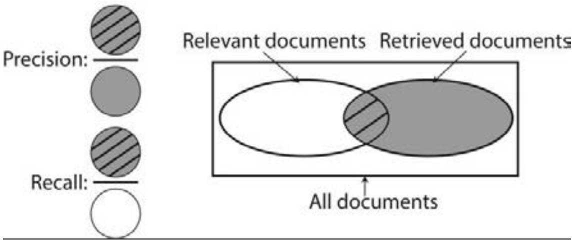

Fig. 1 Graphical representation of recall and precision.

In our case, precision can easily be calculated and monitored as the search proceeds. For the seeker of the bibliographic search, maximum precision means reading only the documents that are interesting for the research.

Since recall cannot be easily calculated, another parameter can be used to evaluate the "completeness" of the information, such as the fraction of questions of the ideal article that have been already addressed by at least one document.

Furthermore, since there is a detailed evaluation of each document, it appears reasonable to introduce another parameter for the evaluation that keeps tracks of the overall value/relevance of the retrieved documents. This parameter can be simply the "sum" of the evaluation of each document.

Therefore, three parameters can be implemented to monitor the performance of the collaborative bibliographic search. Figure 2 shows an example of how the metric can be used to decide the end of a bibliographic search. In this case, we can see how after 24 hours of searches, interesting articles are rarer and completeness is almost constant. Here, it would be possible to concentrate on questions that have not been answered by none of the articles or just stop the searches and start to write the bibliography.

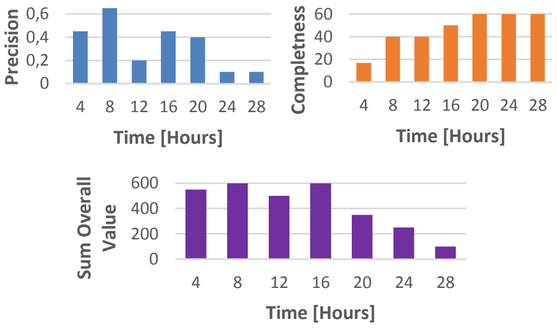

Fig. 2 Monitoring of the bibliographic search

3.5. Bibliographic mapping to communicate

People have always placed a high importance on Data visualization to communicate any result [22].

For the presented collaborative bibliographic search model we propose a graphical representation to illustrate results from one information seeker through a simple template. Publications are sort by originating field on the abscissa axis, and according to the bibliographic search process general timeline on the ordinate axe. Boxes indicate what queries have allowed the documents curations and an arrow between two documents indicates that one has been identified through the other references. Every document is represented as a bubble which size depends on the subjective and objective indexes score.

4. Case Study

Ikos Consulting is an engineering company mainly dedicated to railway technology, collaborating with almost all players in Europe. The railways sector is an important industrial area with intensive research activities. Amongst the most typical topics, there are:

- New materials for wear and stress resistance
- New tracks systems
- Formal methods for increased system safety
- Energy management
- Interoperability, traffic optimization
- Simulation and measurement
- Project management and Life cycle cost optimization

In this context, some case studies have been developed and one of them will be presented.

4.1. Protocol

The case study seized upon the identification of the need for a bibliographic search on a topic defined as: “How formal methods may apply to railway design”. Two researchers performed it in parallel with two different approaches. One research used the presented approach while the second one used an unstructured approach.

- Proposed approach: The researcher and the project’s originator started by defining the Ideal Article together. The questions which the Ideal Publication should answer are the following:

1. What systems have been modeled thanks to formal methods?
2. Do these models share specificities?
3. What is the used language for these modeling?
4. What are the limits of the modeling?
5. Is there an evolution of formal methods use in the railway sector?

Subsequent steps followed the approach disclosed in section 3. The project’s originator re-assessed the selected publication with the same ideal publication criteria.

- Unstructured approach: After an informal interview with the project’s originator, the researcher looked for scientific publications by performing the following query +“railway” +”formal method” in order to acquire a general understanding of the mentioned topic. Listed articles were selected through a title analysis. Selected publications were assessed afterwards by the project originator according to the listed ideal publication criteria.

4.2. Results

Using the two approaches, two sets of articles have been selected. The results from the analysis of these two pools are presented in figure 3 and figure 4.

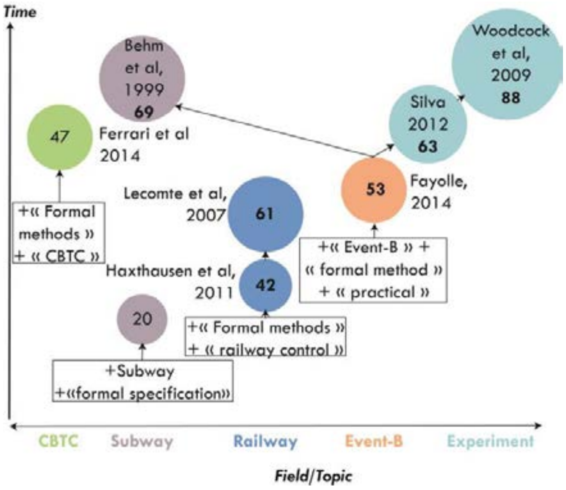

Fig. 3 Results from the proposed approach

The results show that the new entry scores increase over time (from 40 for the 1st query to 57 for the 3rd query). Publications also tends to have increasing scores correlated to the number of iteration. This demonstrates that creating chains of publications tend to increase their value, highlighting the potential contribution of the stigmergy’s concept.

The overall mean for the pool of publication identified through the proposed approach is 59.

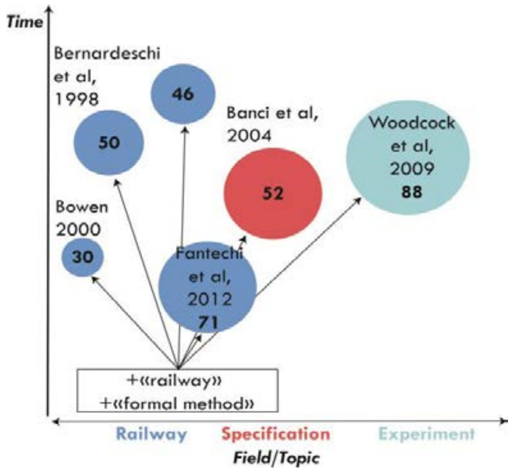

Fig. 4 Results from the unstructured approach

As the approach relies on serendipity, the identified trend of increasing scores does not occur with the unstructured approach. However the required number of iterations to identify the best publication of the pool is shorter (4th publication treated). The overall mean for the pool of publications identified through the proposed approach is 56.

The same post-search analysis than in section 4.3 has been performed on the two pools of publications. Results are presented in Table 3:

Considering the numerical results, the proposed approach provides a slight advantage over the unstructured one with a respective overall mean of 80 and 76.

Table 3. Results of the analysis for the two selected sets of publications.

Questions	% Answered by publications from Pool A (proposed approach)	% Answered by publications from Pool B (unstructured approach)
What systems have been modelled thanks to formal methods?	80	90
Do these models share specificities?	80	90
What is the used language?	90	70
What are the limits of the modelling?	80	60
Is there an evolution of formal methods use in the railway sector?	60	80

Alongside this mean difference, the publications obtained using the two methods appear to be substantially different.

Each article found using the provided approach explains one specific work. It describes a case study, and the use of a method to resolve a problem. The case studies are very detailed. We know the language that has been used, the encountered problems and the limit of the used method.

When using the unstructured approach, the articles are more general. Each article is a state-of-the-art, it points out the general trends in the use of formal methods to specify railway systems.

On a scale of 1 to 10 measuring the usefulness of the approach for someone to take over the bibliographic search, the assessor researcher graded it 8.

5. Conclusion

An important step of any scientific work is the collection of a series of relevant documents that contains valuable information related to the research topic.

Some tools for information retrieval are now used to support this activity and a growing interest has been found in recent publications in collaborative information retrieval.

However, the operative aspects of collaborative bibliography achievement have not been studied in detail. Specifically, the main limits of present methodologies are identified in a limited interaction with the users, a difficult evaluation of the relevance of a document and the difficulties in monitoring the on-going process.

In this paper, the TRIZ Ideal Final ‘Result is used to define ideal documents, which are used to practically define the relevance of a specific document.

Operationally, stigmergy is used as a source of inspiration to align the tasks of different users of the collaborative search. Specifically, seekers are responsible for marking potentially interesting documents, while evaluators read the marked documents to assign subjective indexes of relevance.

An overall index is calculated considering both subjective and objective scores. Based on the value of the overall index, more or less attention is given to the document’s neighborhood, i.e., closely related publications. For instance, when a document with high evaluation is found, also citing and cited documents are included into the analysis.

Communication between subjects is improved with a graphical representation of the evaluated documents. In this way, each subject can have a comprehensive overview on the explored documents.

Thanks to the evaluation, the information retrieval process can be monitored more effectively by plotting performance metrics on a time graph. Monitoring allows a better planning of the research activities, thereby facilitating the identification of the best moment to stop the search and proceed with the writing phase of the bibliography.

References

[1] Greetham, D. Textual scholarship: An introduction; 2013; Routledge.

[2] Francis FC. Bibliography. Encyclopedia Britannica (15th ed.); 1973; Vol. 2: p. 978–981.

[3] Foley C., Smeaton A., “Division of labour and sharing of knowledge for synchronous collaborative information retrieval”. Information processing & management; 2010;46; 6, p. 762-772.

[4] Litvin, S. 2005. “New TRIZ-Based Tool—Function-Oriented Search (FOS).” The TRIZ Journal, August.

[5] Russo D., Montecchi T., Ying L. “Functional-Based Search for Patent Technology Transfer.” In ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference; 2012; p. 529–539.

[6] Hansen P., and Järvelin K.. "Collaborative information retrieval in an information-intensive domain." Information Processing & Management 2005; 41.5; p. 1101-1119.

[7] Hendry D.., Jenkins J. R., McCarthy F. "Collaborative bibliography." Information Processing & Management; 2006;42.3; p. 805-825.

[8] Tomiyama T. "Dealing with Complexity in Design: A Knowledge Point of View." Design Methods for Practice; 2006; p.137-146.

[9] Paletz S., Schunn C. "A Social-Cognitive Framework of Multidisciplinary Team Innovation." Topics in Cognitive Science; 2010; 2: p. 73-95.

[10] Gull C. "Seven years of work on the organization of materials in the special library." American Documentation 7.4, 1956; p. 320-329.

[11] Althuller G., "And Suddenly the Inventor Appeared," Worcester: Technical Innovation Center; 1996.

[12] Bonabeau E., Dorigo M., Theraulaz G. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press; 1999.

[13] Page R., Mitchell S. "Self organization and adaptation in insect societies." PSA: Proceedings of the biennial meeting of the philosophy of science association. Philosophy of Science Association; 1990.

[14] Seeley, T. “When is self-organization used in biological systems?” The Biological Bulletin; 2002; 202: p. 314-318.

[15] Moussaid, M., Garnier, S., Theraulaz, G., Helbing; D. “Collective Information Processing and Pattern Formation in Swarms, Flocks, and Crowds”. Topics in Cognitive Science; 2009; 1: p.469-497.

[16] Wilson E. "The sociogenesis of insect colonies." Science(Washington) 228.4707; 1985; p. 1489-1495.

[17] Grassé, P. "La reconstruction du nid et les coordinations interindividuelles chezBellicositermes natalensis etCubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs." Insectes sociaux 6.1; 1959: p. 41-80.

[18] Dorigo M., Birattari M. "Ant colony optimization." Encyclopedia of machine learning. Springer US; 2010. p.36-39.

[19] Prabhakar B., Dektar K., Gordon D. "The regulation of ant colony foraging activity without spatial information." PLoS computational biology; 2012; (8)8.

[20] Tero A. et al. "Rules for biologically inspired adaptive network design." Science 327.5964; 2010: p. 439-442.

[21] Manning C., Prabhakar R., Schütze H., and others. Introduction to Information Retrieval. Vol. 1. Cambridge university press Cambridge; 2008.

[22] Friendly M., Denis J. "Milestones in the history of thematic cartography, statistical graphics, and data visualization." *U RL http://www. datavis. ca/milestones;* 2001.