Cite-as-you-write

by Kris Jack, Maurizio Sambati, Fabrizio Silvestri, Salvatore Trani, Rossano Venturini

When starting a new research activity, it is essential to study related work. Traditional search engines and dedicated social networks are generally used to search for relevant literature. Current technologies rely on keyword based searches which, however, do not provide the support of a wider context. Cite-as-you-write aims to simplify and shorten this exploratory task: given a verbose description of the problem to be investigated, the system automatically recommends related papers/citations.

Recommender systems for scientific papers have received much attention during the last decade and some novel approaches have been experimented. We propose an innovative contextual search engine tool to address this problem. The tool exploits several aspects of the scientific literature ecosystem including an intriguing social angle derived from data provided by the popular Mendely platform. The objective is to provide pointers to existing literature given a description of the study the researcher is undertaking.

We began by building a baseline method to retrieve literature related to a fragment of text describing the research concept. This consisted of a system taking as input a textual description of the research idea and returning the most similar (according to a similarity measure) papers in the literature. To evaluate the similarity between the query and documents the “cosine similarity” metric was used. This metric calculates how many terms are in common between the query and each document and weights the terms (in query and documents) by an importance score.

We subsequently refined the baseline strategy by adopting a “Learning to Rank” approach. Within this approach the similarity between queries and documents is computed via a metric that is “learned” from samples input to a machine learning algorithm. The main difference between a search engine and Cite-as-you-write consists in how the queries are formulated: search engines are usually optimized to answer keyword-based queries, our system extracts a context from a long description of the research problem provided by the scientist.

The system consists of three modules:

- Crawler, which builds and maintains the repository of scientific literature.
- Indexer, which processes the scientific papers and builds an index on their most representative terms.
- Query processor, a specialized search engine with an advanced ranking algorithm designed for the task of retrieving related work from a detailed context.

The data used to build and evaluate our system consists of about 500 thousand computer science papers including their citations. The data was kindly provided to us by Mendeley.

The index represented under the form of an inverted index contains only the most representative terms for each paper. This trade-off in coverage, keeps down the size of the index. Fortunately. our experiments show that the loss due to reduced coverage is limited, as scientific publications usually focus on a few specific topics represented by a small number of important terms.

Following the typical approach of learning-to-rank-based retrieval systems, the ranking phase consists of two steps:
1. A cosine similarity ranking based on the title and the abstracts of the papers (ie, the baseline method)
2. A ranking function that combines all the features we have collected from the data

The second step works by adopting a technique known as similarity learning, which consists in exploiting sample data to find correlations between a set of features and a target variable. The learning method adopted is a Random Forest ensemble which uses our features based on text similarity, paper publication date, the citation network (i.e. PageRank and absolute number of citations) and the Mendeley environment (ie popularity of papers in user libraries with some normalizations: the importance of the user, defined in the social network as the number of links, and by the number of items in the library, reducing the weight when a user has lots of papers in their library). Random Forest is a very simple but powerful machine learning tool which builds several decision trees from various random samplings of the data. The result is the average of the results returned by each tree. The same strategy as that adopted in democratic voting systems.

Experiments shown in Figure 1 show the improvements on the test data over the baseline system with variations in the number of trees (x-axis).

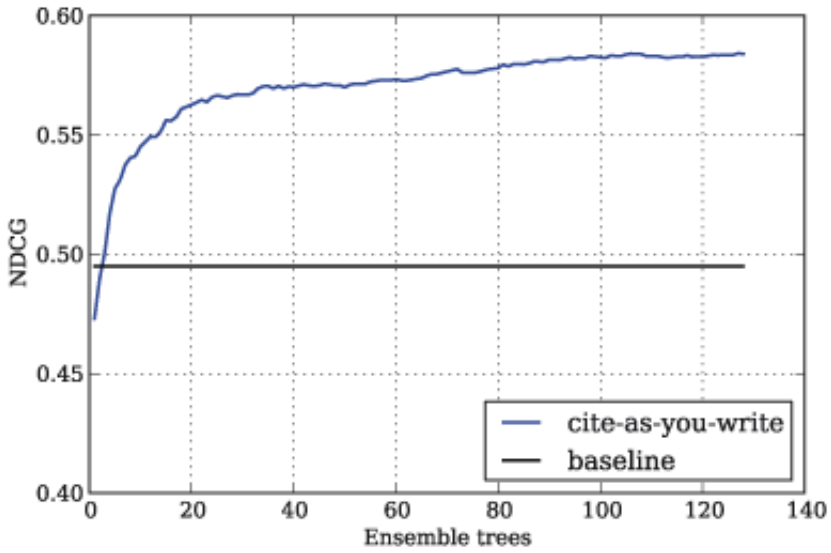

Figure 1: improvements on the test data over the baseline system with variations in the number of trees (x-axis)

We used the normalized discounted accumulative gain (nDCG) evaluation metric to measure the effectiveness of a ranking, ie how near relevant documents were to the top of the result lists [1]. The baseline considered (the black line) is the cosine similarity based ranking. The improvement provided by our tool on the test set with respect to the base line is summarized in Figure 2.

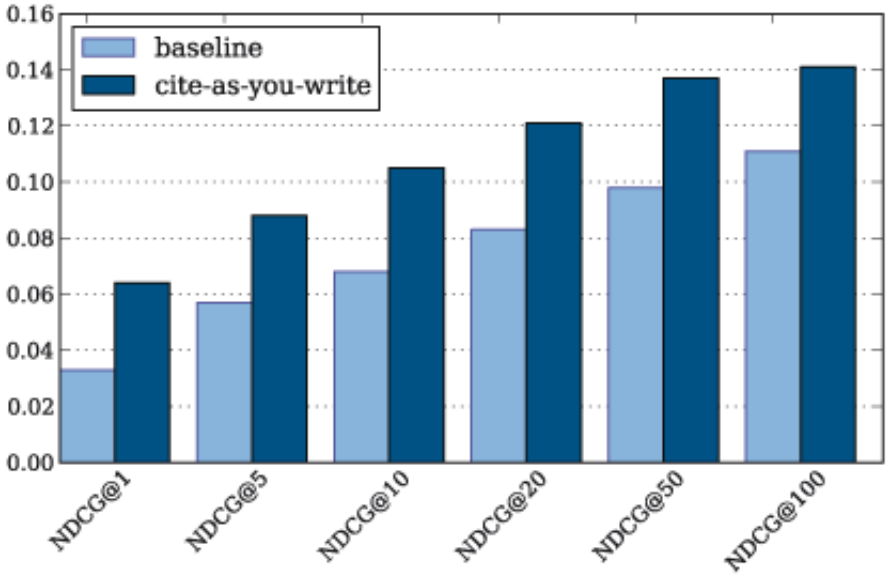

Figure 2: test ranking results

Learning to rank and recommender systems are emerging fields in computer science and have been applied to real world problems in a number of ways. We have used and adapted cutting edge technologies to build a powerful tool for a more natural scientific paper recommendation with respect to traditional search engines. We believe context-driven search can be a better alternative to keyword based search for next generation web.

Links:
Cite-as-you-write (prototype): http://vinello.isti.cnr.it/cite-as-you-write/
Tool used for learning to rank: https://bitbucket.org/duilio/mltool
Mendeley: http://www.mendeley.com/
Random Forests:
http://stat-http://www.berkeley.edu/users/breiman/RandomForests/cc_home.htm

Reference:
[1] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4

(October 2002), 422-446. DOI=10.1145/582415.582418 http://doi.acm.org/10.1145/582415.582418

Please contact:
Maurizio Sambati and Salvatore Trani
ISTI-CNR, Italy
E-mail. maurizio.sambati@gmail.com,
trani.salvatore@gmail.com

Add comment

E-mail (required, but will not display)

Notify me of follow-up comments

Refresh

Send

JComments