The CLAIRE Visual Analytics System for Analysing IR Evaluation Data

Marco Angelini[1], Vanessa Fazzini[1], Nicola Ferro[2], Giuseppe Santucci[1], and Gianmaria Silvello[2]

[1] University of Rome "La Sapienza", Italy, {surname}@diag.uniroma1.it
[2] University of Padua, Italy, {name.surname}@unipd.it

Abstract. In this paper, we describe *Combinatorial visuaL Analytics system for Information Retrieval Evaluation (CLAIRE)*, a *Visual Analytics (VA)* system for exploring and making sense of the performances of a large amount of *Information Retrieval (IR)* systems, in order to quickly and intuitively grasp which system configurations are preferred, what are the contributions of the different components and how these components interact together.

1 Introduction and Context

Information Retrieval (IR) systems are constituted of "pipelines" of components such as stop lists, stemmers and IR models, which are stacked together in order to process both documents and user queries and to match them returning a ranked result list of documents in decreasing order of estimated relevance. Currently, the only viable means to determine the contribution to the system effectiveness of single components is to measure their impact on the overall performances by testing all the different combinations of such components. This leads to a very high number of cases to be considered, making the space of system combinations large and complex to explore.

Besides requiring a great deal of effort and resources to be produced, these combinatorial compositions constitute a challenge when it comes to explore, analyze, and make sense of the experimental results with the goal of understanding how different components contribute to the overall performances and interact together. Indeed, it is typically needed to resort to rather complex statistical tools (e.g. multi-way *ANalysis Of VAriance (ANOVA)* models) requiring a careful experimental design and producing results which call for a considerable extent of expertise to be interpreted [4].

In this paper we present a *Visual Analytics (VA)* system, called *Combinatorial visuaL Analytics system for Information Retrieval Evaluation (CLAIRE)*, which allows for exploring and making sense of the performances of a large amount of IR systems generated from all the possible combinations of components under examination, in order to quickly and intuitively grasp which system

Extended abstract of [1].

IIR 2018, May 28-30, 2018, Rome, Italy. Copyright held by the author(s).

CEUR-WS.org/Vol-2140/paper12.pdf

configurations are preferred, what are the contributions of the different components and how these components interact together. To showcase and evaluate CLAIRE, we developed an extensive set of $612 \times 6 = 3,672$ systems – i.e. the *Grid of Points (GoP)* [3] – arising from the combinatorial composition of several open-source publicly available components such as stop lists, stemmers, and IR models, and run against 6 different public test collections shared by the *Text REtrieval Conference (TREC)* international evaluation initiative.

CLAIRE addresses the necessity to analyse large combinations of system components due to the proliferation of open source IR systems [7] which allow researchers to easily run systematic experiments. Much less attention has been devoted to applying visual analytics techniques to the analysis and exploration of the performances of IR systems; Angelini et al. in [2] dealt with large-scale evaluation campaigns, a context in which evaluators do not have access to the tested systems but they can only examine the final outputs. Other approaches can be found in [5] and in [6]; however, none of these approaches dealt with the inspection of both configurations and measures of IR systems.

2 The CLAIRE system

The CLAIRE system is available at `http://awareserver.dis.uniroma1.it:11768/claire/`. We considered three main components of an IR system: stop list, stemmer, and IR model. We selected a set of alternative implementations of each component and, by using the Terrier v.4.0 open source system, we created a run for each system defined by combining the available components in all possible ways. The selected components are:

- *Stop list*: `nostop`, `indri`, `lucene`, `snowball`, `smart`, `terrier`;
- *Stemmer*: `nolug`, `weakPorter`, `porter`, `snowballPorter`, `krovetz`, `lovins`;
- *Model*: `bb2`, `bm25`, `dfiz`, `dfree`, `dirichletlm`, `dlh`, `dph`, `hiemstralm`, `ifb2`, `inb2`, `inl2`, `inexpb2`, `jskls`, `lemurtfidf`, `lgd`, `pl2`, `tfidf`.

We considered the following standard and shared collections, each track using 50 different topics, *TREC Adhoc tracks* `T07` *and* `T08`, *TREC Web tracks* `T09` *and* `T10`, *TREC Terabyte tracks* `T14` *and* `T15`. Overall, these components define a $6 \times 6 \times 17$ factorial design with a GoP consisting of 612 system runs. They represent nearly all the state-of-the-art components which constitute the common denominator almost always present in any IR system for English retrieval and thus they are a good account of what can be found in many different operational settings.

The design of the system followed agreed solutions in the field of Infovis: the CLAIRE main structure relies on *multiple coordinated views*. CLAIRE comprises the three main areas shown in Figure 1. (1) *Parameters Selection* area, dealing with the exploration coordinates, i.e., collections, stop lists, stemmers, IR models, and measures. (2) the *System Configurations Analysis* area, enabling

`http://www.terrier.org/`

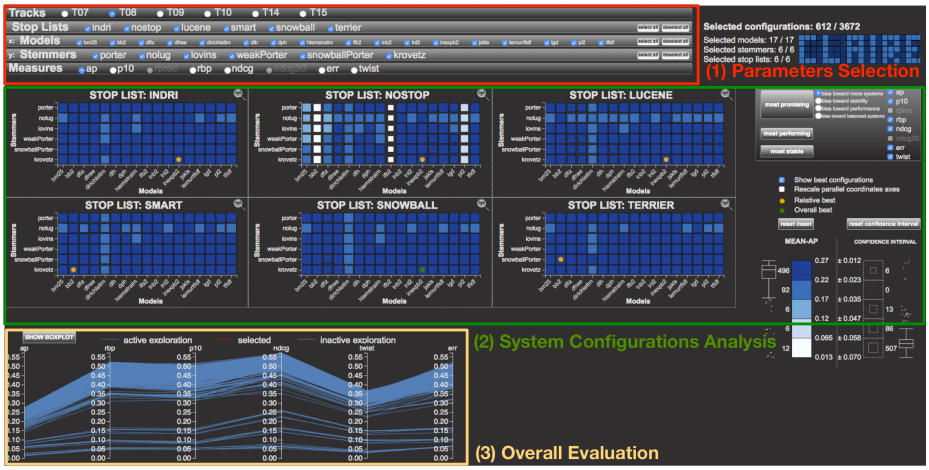

Fig. 1: A comprehensive view of the CLAIRE system. We see a sequence of 6 tiles corresponding to the available stop lists; each tile presents the 17 IR models on the x-axis and the 6 stemmers on the y-axis. The selected track is `T08` and the evaluation measure is AP.

the performance analysis of the system configurations using the actual evaluation measure where each box represents a specific component (stop list in the figure) and the tiles within the box contain all the combinations of the other two components (IR models and stemmers in the figure) with the component in the box. Each tile represents a system configuration, with its color, ranging from white to deep blue, represents a low mean value or a high mean value respectively, averaged over all the topics of a track. The size of the tile represents the confidence interval (small sizes tied to low values), again averaged over all the topics of a track. (3) the *Overall Evaluation* area, where the system configurations performances are evaluated on the complete set of given evaluation measures. Moreover, to automate the selection of relevant subsets of systems, CLAIRE uses the available measures to select clusters of similar systems.

In order to present an example of obtainable evidences from the use of CLAIRE, Figure 2 reports three model tiles corresponding to different visual archetypes: (a) the `bb2` model needs a stop list to function well, (b) the `tfidf` model works better with a stemmer, and (c) the `bm25` model suffers the absence of the stop list and also, even though the effect is less marked, the absence of a stemmer; It is also possible to see that the three models need a stop list and/or a stemmer, but they do not discriminate between different stop lists and stemmers.

3 Conclusions and Future works

We presented a relevant case implying the exploration of almost 1.5M data points corresponding to different performance measures of hundreds of IR systems. To

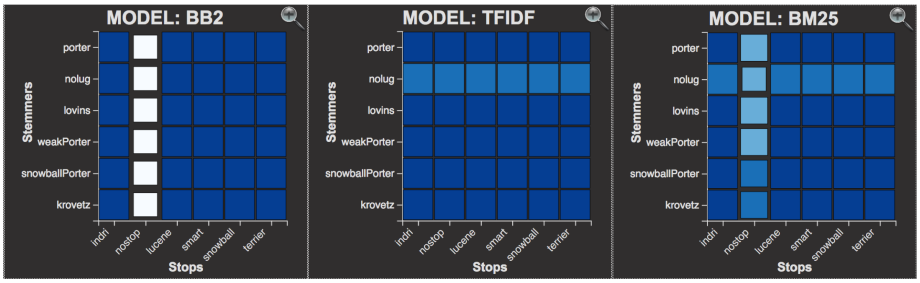

Fig. 2: Three visual archetypes identified from the model tiles: (a) a lighter column shows a problem with a stop list; (b) a lighter row shows a problem with a stemmer; (c) a lighter cross shows a problem with both a stop list and a stemmer.

this end, we developed a novel VA system, CLAIRE, that supports the analysis of a large set of IR systems. As future work, we will extend the CLAIRE system by allowing users to upload their proprietary systems and components and compare them against the standard open-source baselines present in the CLAIRE GoP.

References

1. Angelini, M., Fazzini, V., Ferro, N., Santucci, G., Silvello, G.: CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. Information Processing & Management (2018)
2. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. Journal of Visual Languages & Computing (JVLC) 25(4), 394–413 (August 2014)
3. Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF Pilot Track Overview. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers. pp. 552–565. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany (2010)
4. Ferro, N., Silvello, G.: Toward an Anatomy of IR System Component Performances. Journal of the American Society for Information Science and Technology (JASIST) 69(2), 187–200 (February 2018)
5. Ioannakis, G., A.Koutsoudis, Pratikakis, I., Chamzas, C.: RETRIEVAL - An Online Performance Evaluation Tool for Information Retrieval Methods. IEEE Trans. Multimedia 20(1), 119–127 (2018)
6. Lipani, A., Lupu, M., Hanbury, A.: Visual Pool: A Tool to Visualize and Interact with the Pooling Method. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017). ACM Press, New York, USA (2017)
7. Trotman, A., Clarke, C.L.A., Ounis, I., Culpepper, J.S., Cartright, M.A., Geva, S.: Open Source Information Retrieval: a Report on the SIGIR 2012 Workshop. ACM SIGIR Forum 46(2), 95–101 (December 2012)