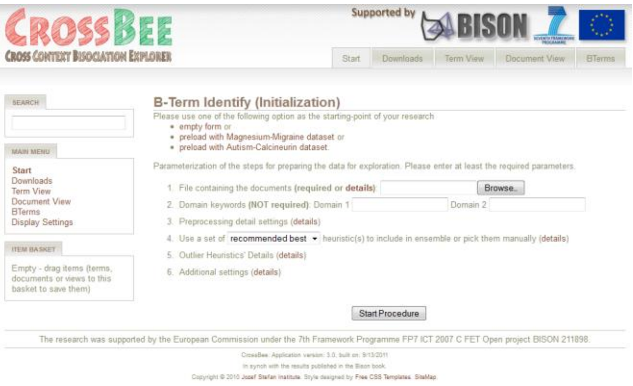
CrossBee: Cross-Context Bisociation Explorer

Matjaž Juršič[1,2], Bojan Cestnik[3,1], Tanja Urbančič[4,1], Nada Lavrač[1,4]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] Temida d.o.o., Ljubljana, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia
{matjaz.jursic, bojan.cestnik, tanja.urbancic, nada.lavrac}@ijs.si

CrossBee is an exploration engine for text mining and cross-context link discovery, implemented as a Web application with a user-friendly interface. The system supports the expert in advanced document exploration supporting document retrieval, analysis and visualization. It enables document retrieval from public databases like PubMed, as well as by querying the Web, followed by document cleaning and filtering through several filtering criteria. Document analysis includes document presentation in terms of statistical and similarity-based properties, topic ontology construction through document clustering. A distinguishing feature of CrossBee is its powerful cross-context and cross-domain document exploration facility and bisociative (Koestler 1964) term discovery aimed at finding potential cross-domain linking terms/concepts. Term ranking based on an ensemble heuristic (Juršič et al. 2012) enables the expert to focus on cross-context links with high potential for cross-context link discovery. CrossBee's document visualization and user interface customization additionally support the expert in finding relevant documents and terms through similarity graph visualization, a color-based domain separation scheme and highlighted top-ranked bisociative terms.

A typical user scenario starts by inputting two sets of documents of interest and by regulating the parameters of the system. The required input is a file with documents from two domains. Each line of the file contains exactly three tab-separated entries: (a) document identification number, (b) domain acronym, and (c) the document text. The other options available to the user include specifying the exact preprocessing options, specifying the base heuristics to be used in the ensemble, specifying outlier documents identified by external outlier detection software, defining the already known bisociative terms (*b* terms), and others. Next, CrossBee starts a computationally very intensive step in which it prepares all the data needed for the fast subsequent exploration phase. During this step the actual text preprocessing, base heuristics, ensemble, bisociation scores and rankings are computed in the way presented in the previous section. This step does not require any user intervention. After this computation, the user is presented with a ranked list of *b* term candidates. The list provides the user with some additional information including the ensemble's individual base heuristics votes and term's domain occurrence statistics in both domains. The user then browses through the list and chooses the term(s) he believes to be promising *b* terms, i.e. terms for finding meaningful connections between the two domains. At this point, the user can start inspecting the actual appearances of the selected term in both domains, using the efficient side-by-side document inspection. In this way, he can verify whether his rationale behind selecting this term.

CrossBee is available at website: http://crossbee.ijs.si/. The system's home page is shown below.



B-Term Identify (Initialization)

Please use one of the following option as the starting-point of your research

- empty form or
- preload with Magnesium-Migraine dataset or
- preload with Autism-Calcineurin dataset.

Parameterization of the steps for preparing the data for exploration. Please enter at least the required parameters.

1. File containing the documents (required or details):
2. Domain keywords (NOT required): Domain 1 Domain 2
3. Preprocessing detail settings (details)
4. Use a set of recommended best heuristic(s) to include in ensemble or pick them manually (details)
5. Outlier Heuristics' Details (details)
6. Additional settings (details)

The research was supported by the European Commission under the 7th Framework Programme FP7 ICT 2007 C FET Open project BISON 211898.

CrossBee: Application version: 3.0, built on: 9/13/2011
In synch with the results published in the Bison book.
Copyright © 2010 Jozef Stefan Institute. Style designed by Free CSS Templates. SiteMap

References

Juršič, M.; Sluban, B.; Cestnik, B.; Grčar, M.; and Lavrač, N. 2012. Bridging concept identification for constructing information networks from text documents. In: Berthold, M.R. ed., *Bisociative Knowledge Discovery*. Springer LNAI 7250 (in press).

Koestler, A. 1964. *The Act of Creation*. New York: MacMillan.