# C-MLink: a Web-based Tool for Transitive Text Mining

Guenter Grohmann [*] and Johannes Stegmann [**]

[*] *guenter.grohmann@charite.de*
Charité - University Medicine Berlin, Campus Benjamin Franklin, Institute of Medical Informatics, Biometry and Epidemiology, D-12203 Berlin (Germany)

[**] *johannes.stegmann@charite.de*
Charité - University Medicine Berlin, Campus Benjamin Franklin, Medical Library, D-12203 Berlin (Germany)

**Introduction**

Transitive text mining tries to establish meaningful links between the main concepts of non-overlapping literature sets. The basic ideas, several examples and an interactive software system offering some help in finding transitive links have been developed and published by Don Swanson (e.g. Swanson, 1986; Swanson 1988; Swanson & Smalheiser, 1997). Hence, "Swanson Linking" (SL) might be used as the appropriate term for this kind of literature-driven hypothesis generation.

The presentation will describe the main features and first experiences of an advanced information retrieval system developed for hypothesis building and "knowledge discovery" tasks which will be implemented in the next future at the authors' institution. The system basically analyses the MeSH terms, i.e. the MEDLINE keywords assigned to the sets of PubMed documents serving as data input (Stegmann & Grohmann, 2003; Srinivasan, 2004).

The name of the system is "C-Mlink" where "C" means "Charité", and "M" may be read as "Missing" or "Mining".

**SL – Basic Issues**

In general, a researcher tries to find solutions for a problem by testing hypotheses. Due to the large amount of published research articles it might happen that the literature dealing with the research problem does not contain all possible ideas serving as input for hypothesis building. However, there may exist "intermediary" concepts bridging the semantic "gap" between the "source" (problem related) literature and "target" literature(s), the latter possibly contributing to problem solving but being not found by conventional retrieval of the source literature. SL tries – by whichever method implemented - to identify possible intermediary and target concepts.

**C-MLink: How It Works**

As a registered user a scientist starts a mining session within a standard browser. A typical session consists of five main steps which may be repeated with changed document sources as often as needed. First, the user has to conduct a PubMed search for problem-related literature, e.g. dealing with a disease, and to download the resulting documents in MEDLINE- or XML-format. The next step comprises the upload of the downloaded set to the C-Mlink server and - after adjusting some runtime program variables (like thresholds for term occurrence, link strength, clustersize) – the launch of the MeSH extraction and analysis tools. C-MLink then provides the results of the keyword extraction and analysis tasks in textual (e.g. keyword-cluster-tables) and graphical form (cluster centrality/density diagrams) and tries to give some hints – on the basis of cluster characteristics – which terms are most promising to serve as intermediate concepts. Then, the user has to decide which of the intermediary concepts should be followed, retrieve and upload the appropriate literature to the C-MLink server which in turn gives back possible target terms on the basis of their cluster characteristics (Stegmann & Grohmann, 2003).

The user gets additional support in investigating the respective knowledge domain through a kind of (hyperlink) navigator, which provides detailed information of interesting terms and clusters (corresponding titles, abstracts, cluster locations).

**Software Design**

C-MLink uses the Java-2 Enterprise Edition (J2SEE) which defines a standard for implementation, configuration, distribution and deployment of Enterprise applications. This framework is based on the component-model (client, web, program component) and the java-api, particularly jdbc,servlets,beans. The J2EE-model is web-based. Clients access the middleware via HTTP where they use HTML or XML. They communicate primarily with web containers which execute servlets, java server pages (JSP) or beans. As a guide for the design the model-view-controler paradigma (MVC) is utilised. It divides a complex application into logical units: the presentation layer (the views/front-ends, realized by server pages), the controler object (e.g. servlets, handling GET or POST requests, event handler), and the model object (beans, pure java) which handles the core application logic. The main benefits of the MVC are the strict distinction

between design and application logic, its good maintainablility and scalability.

**Implementation**

As a servlet container C-MLink utilizes Tomcat (cf. Apache Foundation), which is SUN Microsystems' reference implementation for the J2EE-Environment. The C-MLink controler calls the core text mining modules (in the model layer). These are pure java objects like: *Doc_Analysis_Get_Term, Doc_Analysis_Calculate, Doc_Analysis_Cluster*.

The java class objects implement a set of methods like *calculate_strength*, *calculate_density, calculate_centrality*. The system is able to accept the upload-input either in plain ASCII or in XML format. The data transfer between the views (JSP), the application, and the data-persistance-layer (backend = MySQL) is managed by a bunch of beans with GETer and SETer-methods like *setUpload_file(), getUpload_file(), setField_of_interest(), getField_of_interest*. Visualisation will be made possible via invoking appropriate software packages.

**Development and Runtime Environment**

Open-source software is used for all relevant parts of the C-Mlink service:

Java 2 Development Kit Enterprise Edition 1.4; Java Servlet API 1.3; Java JDBC API; Upload Beans from Jason Hunter's O'Reilly Servlet package; Eclipse 2.1 (IDE for java-source development); Tomcat 5.1 (Web-Server); MySQL Ver. 4.1 (Database).

C-Mlink runs under SUSE Linux 9.0 on a Pentium IV PC.

**Planned Add-on**

It is planned to implement natural language processing tools for full-text analysis (titles, abstracts) of PubMed documents. In addition, a tool will be implemented which allows screening of terms in the neighbourhood of relevant source (and intermediate) MeSH terms after eigenvalue decomposition of document-by-MeSH-term matrices generated from the uploaded literature sets (Stegmann & Grohmann, 2005).

**Acknowledgements**

Our work is currently supported by the Deutsche Forschungsgemeinschaft, grant no.LIS-4-54281.

**References**

Srinivasan, P. (2004). Text Mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55, 396-413.

Stegmann, J. & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. *Scientometrics,5*6,111-135.

Stegmann, J. & Grohmann, G. (2005). Factor analytic approach to transitive text mining using PubMed keywords. *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics* (ISSI 2005).

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30, 7-18.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31, 526-557.

Swanson, D.R. & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91, 183-203..