Superconductor Research Papers Clustering using Annotated Information

Sae Dieb[1*], Luca Foppiano[1], Kensei Terashima[2], Pedro Baptista de Castro[2], Masaharu Yoshioka[3],Yoshihiko Takano[2] and Masashi Ishii[1]

[1]MaDIS, National Institute for Materials Science, Tsukuba, Japan. [2]MANA, National Institute for Materials Science, Tsukuba, Japan [3]IST, Hokkaido University, Sapporo, Japan.

[*]E-mail: DIEB.Sae@nims.go.jp

In this presentation, we report an on-going work to cluster research papers on superconductor domain focusing on a specific category(s) of information using a weighing mechanism for different categories of information. We aim to facilitate finding relevant information centered around the interest of the superconductor researchers, such as the class of the superconducting material or the critical temperature measurement method. Clustering research papers using the general content of the paper might not be efficient for researches interested in one or more information categories.

In this work, we use annotated papers from the superconductor domain from the SuperMat corpus [1]. SuperMat corpus consists of research papers annotated with linked 6 information categories related to superconductors development. We compare the clustering results between 2 cases, in the first, we use certain information categories from SuperMat corpus, and the second, we use the non-annotated papers of the same corpus. The comparison is measured by internal evaluations metrics.

Papers were preprocessed to reduce noise by removing English stop words, punctuations , numeric values and physical quantities. In the next step, each paper is transferred into a vector using bag of words approach. For the non-annotated paper, it is composed of one vector represents the vocabulary that exist in the paper and their frequencies. On the other hand, the annotated paper is transformed into multiple vectors, each vector contains a bag-of-words representation of all annotations under certain information category. We consider 4 information categories: class of superconducting material, materials, measurement method, and the non-annotated tokens named as ”other”.

We used agglomerative clustering method with the euclidean distance and ward method to determine papers similarity to each other. Figure 1 compares the clustering results for non-annotated version of the corpus (a) and the annotated version focusing on the measurement method information category(b). A rule-based system was used to categorize the measurement method annotation into 4 groups as follows: magnetization, calculation, resistivity and specific heat. Silhouette score showed higher clustering quality for annotated version of the papers.

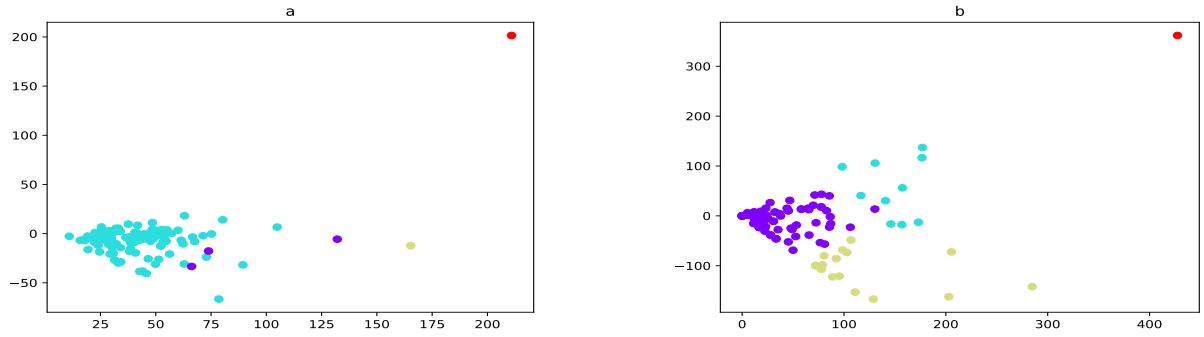

Figure 1: Truncated SVD projection for the agglomerative clustering result for (a) non-annotated version of the corpus, and (b) the annotated version using the following weights array [1,1,10,0] corresponding to [class, material, measurement method, non-annotated text]

[1]L. Foppiano, S. Dieb, A. Suzuki, et. al. SuperMat: Construction of a linked annotated dataset from superconductors-related publications, preprint