DADC for Web Search

There are a variety of clustering techniques that can be used to improve the quality of results in one of the largest topics in text analysis: web search. One such approach, called Dynamic Agglomerative-Divisive Clustering (DADC), analyzes clickthrough data to group various objects that support web search in two phases. As new clickthrough data is inputted into the web search, both the agglomerative and divisive phases are used to form effective clusters. In this paper I describe and review the model and its applications.

The foundation of this algorithm lies in a graph model called a Community Clickthrough Model (CCM). When a user uses web search, the results they click on are used as indicators of the user's intention, which in turn allow the search engine to suggest better queries in the future to support the user. CCM replaces simple click data with a "concept" which is embedded in the clicked-on document itself. The addition of "concepts" to CCM is an attempt to control for one of the weaknesses of many clickthrough clustering methods: content ignorance. CCM is a tripartite graph because it represents a relationship between the user, his/her query, and the concept of the document which has been clicked on.

In CCM, a concept is identified using support formula as a measure of "interestingness'. Web snippets are used as the keyword in the support formula. However, we could modify the CCM formula to incorporate one of the many other techniques for finding frequent itemsets to identify concepts for use in the tripartite graph.

The DADC algorithm updates the CCM graph in two phases: the agglomerate phase and the divisive phase. The agglomerative phase merges similar clusters iteratively every time new data is added and the algorithm updates the graph. The divisive phase splits clusters into smaller ones

in order to prevent them from growing too large when new concepts are added to the graph. Clickthrough data is converted into CCM as a precursor to both phases.

In the agglomerative phase, there are a few assumptions made which are related to similarity of each aspect of the tripartite graph: users, queries, and concepts. The first is that two users are similar if they send similar queries to the search engine and click on results with similar concepts. The second is that two queries are similar if they're submitted by similar users for similar concepts. The third is that two concepts are similar if they are opened by similar users looking for similar queries. A weight vector is used to change the weights of the edges between the user, query, and concept nodes depending on the similarity measures. Clusters are iteratively merged until no new clusters can be created.

In the subsequent divisive phase, hierarchical clustering is used to split large clusters so they don't grow too large. That is, it continues to split clusters until the clusters are not able to be split any longer. Because it's important to learn the minimum number of observations needed for this phase, the Hoeffding bound is applied to the distance equation used to determine if a cluster should be split or not. Together with the agglomerative phase, this phase ensures that the clusters formed during web search create optimal search results.

While DADC has been used to explain how a web search engine like Google can be made more effective, it could also be applied to other types of recommender engines. In fact, any query based recommender engine can benefit from the agglomerative-divisive iterative clustering process. For example, online retailers like Amazon could use this algorithm to suggest the correct product to the shopper. In this case, the query would be the item the shopper is looking for, the user would be the shopper, and finding the concept would involve parsing the description

of the item links the user clicks on while evaluating his/her options. This data would get converted into a tripartite graph to be used for the algorithm as above.

The DADC algorithm is a unique method for getting useful search results out of a web search. The algorithm's clickthrough data to graph conversion, followed by its two part iterative algorithm, creates more accurate clusters than common clustering algorithms. Going forward, this algorithm can be used in combination with other methods, like collaborative filtering, to achieve the best search results.

References

Leung K.WT., Lee D.L. (2010) Dynamic Agglomerative-Divisive Clustering of Clickthrough Data for Collaborative Web Search. In: Kitagawa H., Ishikawa Y., Li Q., Watanabe C. (eds) Database Systems for Advanced Applications. DASFAA 2010. Lecture Notes in Computer Science, vol 5981. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12026-8_48