Semantic Association for Literature Based Discovery

Sam Henry, M.S. and Bridget McInnes, Ph.D.
Virginia Commonwealth University, Richmond, VA

Introduction

This research explores the effectiveness of association measures at ranking the output terms of a literature based discovery (LBD) system. The amount of biomedical text published is growing exponentially and researchers are finding it increasingly difficult to keep up with new findings, even inside their area of expertise. LBD attempts to address this situation by automatically uncovering new, potentially meaningful relations between terms that could lead to new discoveries. A core challenge of LBD is that systems generate more potential discoveries than can be analyzed by a human, therefore effective ranking measures are essential.

Method

Using the traditional ABC co-occurrence model of LBD we begin with a start term, *A*, from which *A implies B* relationships are found via co-occurrences in text. Using the generated *B* terms, *B implies C* relationships are found. From these, *therefore A implies C* relationships are inferred to produce a list of output terms. This method generates hundreds or even thousands of *C* terms, many of which are uninformative and uninteresting, therefore ranking is critical. We evaluate the ranking procedures of: Average Minimum Weight, where the minimum value between each *A* to *B* and *B* to *C* is averaged over all *B* terms for each *C* term; Maximum *B* to *C*, the maximum *B* to *C* value over all *B* terms for each *C* term; Linking Term Count (LTC), where the count of unique *B* terms for each *C*.

Generally, the co-occurrence frequency between terms is used as the “value” in these procedures; we modify them by by replacing the co-occurrence frequency values with association measure values. Association measures quantify the likelihood of two terms occurring together in text versus by chance. Both the terms’ individual occurrence frequencies, and their mutual co-occurrence frequencies are taken into account. This study presents a comprehensive comparison between association measures, including Log Likelihood Ratio, Left Tailed Fisher Test, Pearsons Chi Squared, Dice Coefficient, Odds Ratio, Mutual Information, Jaccard Measure, and Phi Coefficient. We use data prior to year 2000 as the training set, and data published after year 2000 as the test set. We use the 2015 MetaMapped MEDLINE baseline as our dataset, and apply concept filtering to restrict *B* and *C* terms to specific UMLS semantic types.

Evaluation and Results

We evaluate the ranking methods with discovery replication and time slicing evaluation. Discovery replication reproduces previous discoveries, and the rank of the *C* term of interest is reported. The higher the rank, the better the system. We replicate three benchmark discoveries, Raynaud’s Disease - Fish Oil, Migraine - Magnesium, and Somatomedin C - Arginine. For time slicing evaluation, we divide the dataset into testing and training portions. The training portion is used to generate *C* terms, and the test portion is used to simulate *to be discovered* knowledge. 100 *A* terms are randomly chosen from the training set, and *C* terms are generated. Using the *C* terms generated, and the test set as a gold standard, precision and recall curves, Mean Average Precision, and precision at k (precision using only the top k ranked terms) are calculated. Table 1 shows the results of discovery replication. The results show the efficacy of using association measures, but more analysis is required. Further results and analysis are shown on the poster.

			Average Minimum Weight								
	Total	LTC	Frequency	Log Likelihood	Left Fisher	Chi Squared	Dice	Odds	Mutual Info	Jaccard	Phi
Fish Oil	51931	2544	2345	15025	28396	6362	2086	5460	15025	2086	6362
Magnesium	56243	192	192	31970	51915	22973	2408	30837	31970	2408	22973
Arginine	87671	17	38	2020	35133	29422	1352	30766	2020	1352	29422
	Total	LTC	Maximum *B* to *C*								
Fish Oil	51931	2544	2345	1487	45759	4966	3220	18791	1487	3220	4966
Magnesium	56243	192	192	1632	56243	40262	30256	40557	1632	30256	40262
Arginine	87671	17	38	70	86530	810	393	86523	70	393	810

Table 1: Discovery Replication Results for three discoveries and ranking methods.