Linking term association: ranking implicit terms in literature based discovery

Sam Henry, M.S.[1]
[1]Virginia Commonwealth University, Richmond, VA

Aims and Objectives

This research explores the effectiveness of implicit term ranking methods for literature based discovery (LBD). We introduce linking term association (LTA) measures, a modification of co-occurrence association measures designed specifically for ranking implicit information generated by LBD systems. Results are compared to baselines of linking term frequency (LTF) and linking term count (LTC). Three historic discoveries are replicated, and time slicing analysis is performed.

Justification for the Research Topic

The amount of published biomedical text is growing exponentially and researchers find it increasingly difficult to keep up with new findings, even inside their area of expertise. Researchers are increasingly specialized, and overlapping research findings often go unnoticed. LBD attempts to address this problem by automatically uncovering new, potentially meaningful relations between terms that could lead to new discoveries. These inferred discoveries are implicit in text, but not explicitly stated. LBD systems typically generate more potential discoveries (represented as terms) than can be analyzed by a human, therefore ranking these implicit terms is critical.

Many proposed implicit ranking methods are simple modifications of explicit term ranking methods or based purely on frequency (such as LTC and LTF). LTC ranks target terms based on the number of unique linking terms it shares with the start term. LTF ranks target terms based on the sum of co-occurrences of linking terms it shares with the start term. LTC and LTF are purely based on frequency and both are biased to weight uninformative terms that co-occur with many terms (e.g. "Patient", "DNA") more strongly. Association measures use both single term occurrence counts and coupled-term co-occurrence counts to quantify the association between two terms, reducing this frequently occurring term bias. By modifying association measures to use linking term counts as inputs rather than co-occurrence counts we hope to develop more effective implicit term ranking measures for LBD systems.

Research Questions

1. Can we reduce frequently occurring term bias of LTC and take advantage of its empirically reported good performance[1] for the task of ranking implicit terms generated by an LBD system?
2. What is the best way to modify co-occurrence based association measures to rank implicit terms, which by our definition do not co-occur with each other?
3. Does using the count of co-occurrence of each co-occurring term improve results over a using unique co-occurring term counts?
4. What is the best association measure for this task: Mutual Information, Log Likelihood Ratio, Left Tailed Fisher Test, Phi Coefficient, Pearson's Chi Squared, Dice Coefficient, Jaccard Measure, or Odds Ratio?

Research Methodology

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, again via co-occurrences in text. Next, therefore A implies C relationships are inferred to produce a list of target terms. We rank this list of target terms using the baseline techniques of LTC and LTF, and our proposed method, LTA, which is a modification of co-occurrence based association measures that uses linking term counts as input.

Typically, association measures use explicit co-occurrence information as input to quantify the association between two terms. These input values are: *N11* – the count of term 1 followed by term 2, *N1P* – the count of term 1 followed by any term, *NP1* – the count of term 2 preceded by any term, and *NPP* – the count of any term followed by any term (the total number of co-occurrences). Using these four values, statistical information using term occurrence and co-occurrence counts are calculated, and the association between a term pair is quantified. We modify the input values of these measures to use linking term count information rather than co-occurrence information. We redefine: *N11* –

the number of linking terms shared by term 1 and term 2, *N1P* – the count of term 1's linking terms, *NP1* – the count of term 2's linking terms, and *NPP* – the vocabulary size (total number of possible linking terms). This provides us with an intuitive method for ranking implicit terms generated by an LBD system that takes advantage of the empirically validated strength of LTC, and theoretical strength of association measures.

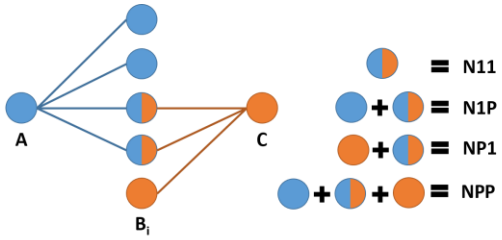

Figure 1: a pictorial representation of the input values of LTA measures

Our system uses the titles and abstracts from the years 1975 to 2015 of the 2015 MetaMapped MEDLINE baseline. Prior to 1975 only about 2% of the citations contain abstracts. We collect co-occurrence information within a window size of 8 on each side of the term, ignore word order, and eliminate all co-occurrences that occur five times or less throughout the entire corpus. We apply concept filtering to restrict B and C terms to specific UMLS semantic groups[1]. Evaluation is performed by replicating three historic discoveries (Raynaud's Disease - Fish Oil, Migraine - Magnesium, and Somatomedin C – Arginine). We report the ranks of target terms of interest; a higher rank indicates better performance. We also perform time slicing evaluation[1], for which we use a cutoff date of January 1, 2000, and divide the corpus into the pre-cutoff portion, containing data prior to the cutoff date (1975-1999), and the post-cutoff portion containing data after the cutoff date (2000-2015). Any co-occurrence present in the post-cutoff portion and absent from the pre-cutoff portion is deemed to be a new discovery, and forms a set of gold standard terms. We generate predicted discovery terms on the pre-cutoff dataset, and report precision and recall curves, mean average precision, precision at k, and frequency at k averaged over 100 randomly generated starting disease terms present in any citations from the year 1999.

Research Results to Date

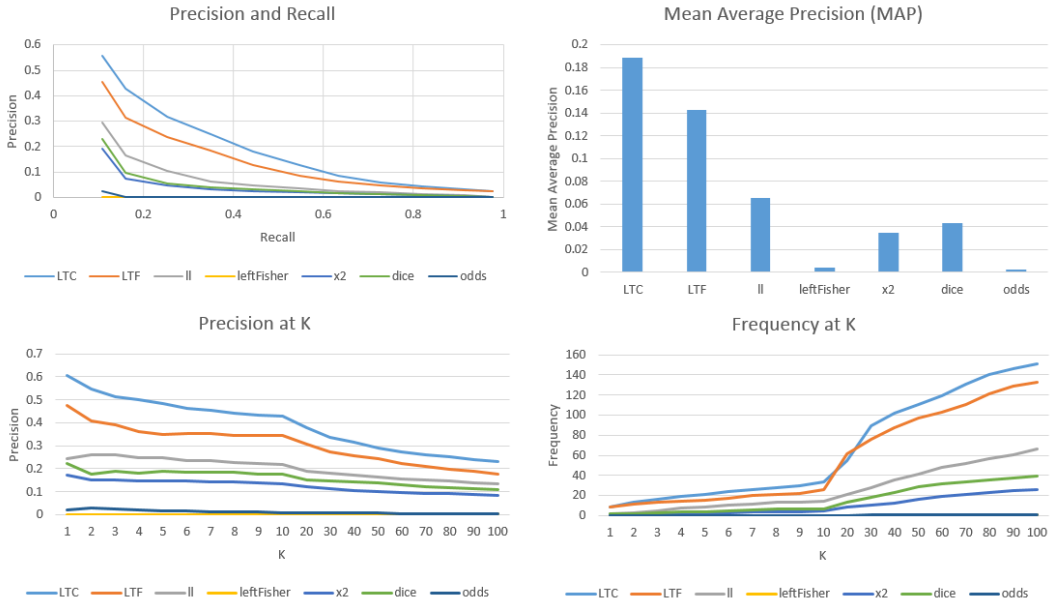

Figure 2. Time slicing results for baseline measures (LTC, LTF) and LTA using different association measures

References

1. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. Journal of biomedical informatics. 2009 Aug 31;42(4):633-43