Evaluation of the Tanaka-Iwasaki-algorithm for word clustering

Mats Mattsson
p02mm@efd.lth.se

Jonas Åström
p02jas@efd.lth.se

Abstract

We have implemented and tested the algorithm described in (Tanaka-Ishii and Iwasaki, 1997).

It is about clustering words based on the co-occurence graph by using transitivity.

We find similiar, but less exact, results. However we have been unable to test the algorithm on a corpus of the same size.

1 Introduction

1.1 Equality relation

A relation can be represented as a graph where vertices a and b are said to be related if there is an edge from a to b. It can be written aRb.

An equality relation (R) is reflective ($aRa, \forall a$), symmetric ($aRb \Rightarrow bRa$) and transitive ($aRb, bRc \Rightarrow aRc$).

1.2 Co-occurrence graph

A graph can be formed from words that co-occur in a corpus. Words are represented as vertices. An edge between two vertices indicate that they co-occur.

This graph can be viewed as an equality relation. Partitioning the graph would give groups of words connected to one topic. Such groups can be used for construction and validation of a thesaurus and clustering of documents.

Both reflectivity and symmetry are guaranteed in the co-occurence graph. Transitivity is usually not present.

1.3 Loosening constraints for subgraph extraction

We loosen the requirement of transitivity for the subgraph. I.e it no longer needs to be a complete graph. Instead of an edge between any two vertices, we only require each vertex in the subgraph to be a part of a complete graph of four vertices. E.g. figure 1.

In (Tanaka-Ishii and Iwasaki, 1997) more theory around the loosening of constraints is discussed.

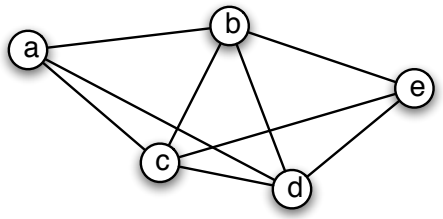

Figure 1: To be a transitive graph an additional edge between vertices a and e is required. After loosening the constraints this graph will be considered transitive.

1.4 Algorithm for clustering

We extract a subgraph A from the co-occurance graph G.

Step 1 Starting from edge e. Put a triangle graph including e into A.

Step 2 For a branch $e' \in A$: If there exists nodes $v \in G$ and $v' \in G$ both forming a triangle with e' and connected to each other, put v and all edges connected to v into A.

Step 3 Repeat step 2 until A cannot be extended any more.

By starting from every triangle in G we will find all subgraphs.

By limiting our output to maximal subgraphs we only have to start from edges not already included in previously calculated subgraph. Some extracted graphs may be parts of others so this needs to be checked.

2 Co-occurance measure

We use the notion of mutual information similar to (Church and Hanks, 1990), which is used in (Tanaka-Ishii and Iwasaki, 1997). Our co-frequency measure is symmetrical and we also

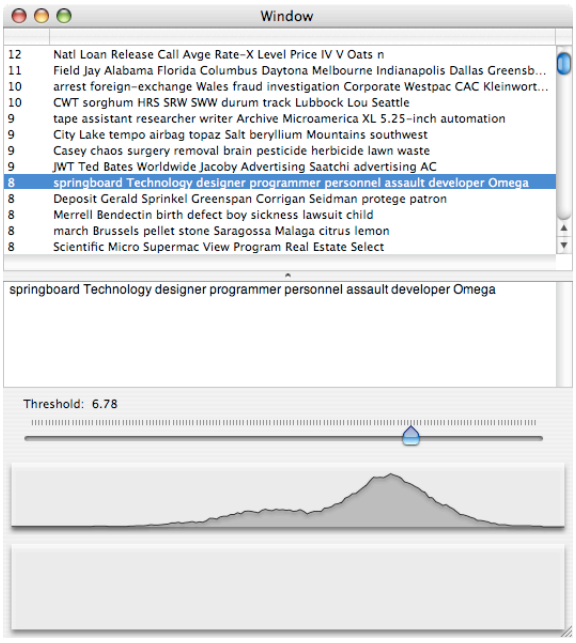

Figure 2: Screenshot of the resuts presented by our implementation. The graph displays the number of clusters for different co-occurance thresholds.

use a finite weighted window

$$w_i = \exp(-\alpha|i| - \beta i^2).$$

As most texts change the subject between paragraphs we add extra distance between them, i.e. the distance between the last word of the previous paragraph and the first word of the current is 7 instead of 1.

To form the graph we set a threshold for the mutual information between two words and say they co-occur when it is above the given threshold.

3 Implementation

We have implemented the algorithm in Objective-C++ as a Cocoa application for Mac OS X. See figure 2 for a screenshot.

3.1 TreeTagger

We use TreeTagger (University of Stuttgart, 2005) to mark the part of speech each word has and find the lemma (e.g. am → be) for words.

We only consider nouns when running the algorithm.

4 Results

As corpora we have used different texts collected from the internet and a part of Reuters-21578 from from the Reuters newswire 1987.

We have only been able to use the algorithm on about 4% of the 15MByte Reuters corpus. Example of the largest clusters:

12 Natl Loan Release Call Avge Rate-X Level Price IV V Oats n

11 (Cities, States) Field Jay Alabama Florida Columbus Daytona Melbourne Indianapolis Dallas Greensboro Jacksonville

10 (Economic crime) arrest foreign-exchange Wales fraud investigation Corporate Westpac CAC Kleinwort Benson

10 CWT sorghum HRS SRW SWW durum track Lubbock Lou Seattle

9 (Writing) tape assistant researcher writer Archive Microamerica XL 5.25-inch automation

9 (Mining) City Lake tempo airbag topaz Salt beryllium Mountains southwest

9 (Cultivation) Casey chaos surgery removal brain pesticide herbicide lawn waste

9 JWT Ted Bates Worldwide Jacoby Advertising Saatchi advertising AC

8 (Designer) springboard Technology designer programmer personnel assault developer Omega

8 Deposit Gerald Sprinkel Greenspan Corrigan Seidman protege patron

8 (Children's disease Merrell Bendectin birth defect boy sickness lawsuit child

8 march Brussels pellet stone Saragossa Malaga citrus lemon

8 Scientific Micro Supermac View Program Real Estate Select

8 Governor Exchequer Nigel Lawson Geoffrey Howe Robin Leigh-Pemberton

8 (Iranian army) attack Iranian Army Revolutionary guard Corps Third commander

About half of the groups have a possible topic, even though there are some noise present.

5 Discussion

In (Tanaka-Ishii and Iwasaki, 1997) a 30MByte corpus from Wall Street Journal is used. They achieve good results for 39 clusters of sizes from 8 to 105 words. There are some noise present but much less than we have encountered.

As the co-occurrence measure by mutual information is more noise resistant for large corpora this may explain the difference between our results and those in the original article.

The running time of our implementation is between 5 and 30 seconds for one clustering of 800kBytes, depending on the mutual information threshold for generating the co-occurance graph.

We have not made much effort to analyze the time complexity of the algorithm. The running time growths worse than linear and probably at least quadratic with the input size.

References

K.W. Church and P. Hanks. 1990. Word associaton norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kumiko Tanaka-Ishii and Hideya Iwasaki. 1997. Clustering co-occurrence graph based on transitivity. *5th Workshop on Very Large Corpora*, pages 91–100.

Institute for Computational Linguistics University of Stuttgart. 2005. Treetagger. *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html*.