

## Discovering Latent Relations of Concepts by Graph Mining Approaches

Marco Büchler

eAQUA Project

Natural Language Processing Group,  
Institute of Mathematics and Computer Science,  
University of Leipzig, Germany  
[mbuechler@eaqua.net](mailto:mbuechler@eaqua.net)

**Abstract:** In the eAQUA project a lot of bridges between Humanities and Computer Science will be built. Whilst the detection of citation or suggestion of words in the process of text completion, the computation of association networks of concepts are an important field.

The computation of association networks is an area of research which has been widely studied. Scientifically, such a network is formalized as a graph  $G=(V,E)$  consisting of a set  $V$  of vertices and a set  $E=V \times V$  of edges between elements of  $V$ . The set  $V$  can be understood as the words of a corpus' vocabulary. Edges between words are computed by co-occurrence analysis. Based on this, a knowledge landscape can be extracted.

In figure 1 two visualisations of graphs are shown. Both graphs are semantic graphs modelled after symmetric sentence based co-occurrences. The left picture displays the context of *Dublin* which can be extracted from an English text corpus. The right example represents two meanings or usages of *space*: mathematics and astronomy.

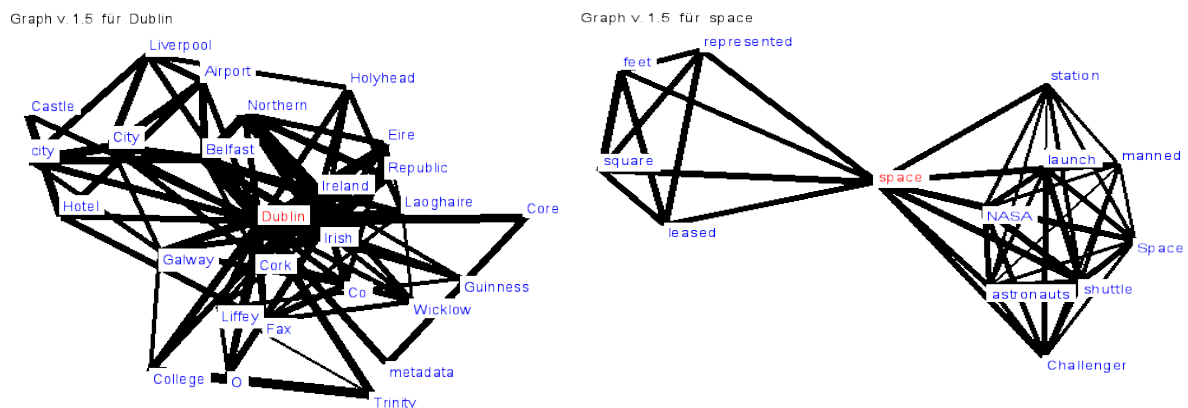


Fig. 1: left: Co-occurrence graph of *Dublin*. *Dublin* is embedded in a single context. Right: Co-occurrence graph of *space* having two different contexts.

Currently, there exists several measures to compute an edge's significance such as the log-likelihood ratio, the Poisson measure or the mutual information as well as the  $\chi^2$  test. However, in Humanities' applications it is important to make a difference between

significance and relevance. While significance computes a mathematical score to measure an association strength, relevance can't be measured by statistics. Relevance depends on what users consider to be important. User's relevance can be clustered by different user groups. That's why students are interested in more obvious relations to explore and learn the research field. They are preferring measures like the log-likelihood ratio or the Poisson measure for their studies. However, a university professor knows the obvious things and is interested in research on rarer or latent associations. Rare events can be found by measures like the mutual information as well as the  $\text{Chi}^2$ .

The main focus of this paper is to identify latent relations. This special kind of relation indicates a non-obvious co-occurrences of two concepts in texts since both words would be expected in a different contextual environment (see figure 1). A very well known example is the relation between *nappies* and *beer* which were bought together in an American supermarket [1]. The reason for that is that the husband was sent to the supermarket to buy nappies. Additionally, he bought a six pack beer for himself. For marketing aspects a relation like that is quite important. Discovering such relations is quite difficult for statistics since they don't occur quite often. For Humanities an approach like that is quite important since the obvious relations are already well-known.

Within the eAQUA project an approach was developed to mine such latent relations. Scientifically, this approach discovers relations which are not intuitive in the first moment (apparent at first glance) and according to this they will be selected. The underlying theory of Selective. Perception will be practiced by every human being to remove unimportant information. Unfortunately, information like that are typically read over by scientists in manual work. Having an objective model, latent relation can be found by the computer quantitatively and after that evaluated qualitatively by Classicists.

Regarding to the introductory discussion of the relation between significance and relevance, this approach builds up a co-occurrence graph of nouns. Based on this, graph relations are not scored by a statistical measure but by the co-occurrence's environment. Typically, words are embedded in a relatively static context. This approach however, investigates and scores relations without this property and selects relations between words by their contextual mutual exclusion.

Within the Chicago Colloquium the technical realisation and efficiency of the above described approach will be explained and shown on several examples based on Ancient Greek texts. Additionally, those results can be clustered by several interests. Finally, the scope to Ancient as well as to Modern texts will be given.

Reference:

[1] <http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html>