**Optimization of References & Normalized and Graphical processing**

**5140309079 高天昊**

# Abstract:

Now, when we are writing thesis, we will select proper references to consult. Publication repositories contain an abundance of information about the evolution of scientific research areas. We address the problem of creating a visualization of a research area that describes the flow of topics between papers, quantifies the impact that papers have on each other, and helps to identify key contributions. To achieve this, I use three ways to integration the evaluation of a paper. Thus, we can come to a result that reflects the true condition of one paper. And then we can analysis the value of the paper and reflect the connection of papers through graph. At the same time, we can also compare two similar fields through this way.

# 1.Introduction :

When we are writing or reading papers of one field, we need to get a quick overview about the research area. There may be a huge amount of papers about it and we need to select the most meaningful ones to consult. This can be provided by a meaningful visualization of the citation network that spins around some given publications.

For example, consider researchers who read on a new topic with some references as starting points. The questions occur which other papers describe key contributions, how the various contributions relate to one another, and how the topic evolved over the past.

This report states the problem we are facing and gives us a reasonable solution, followed by a discussion of related work. And then give us a sample about this problem's solution by a visual graph. The following sections states the solution in detail. It tells you the thinking of solving the problems and then states the procedures of the solution. And I would also like to introduce the way I used to normalize the papers from two fields.

# 2.Evaluation Methodology

I will introduce 3 ways to evaluate the value of the papers and the references among papers and then we draw the result by a graph.

## 2.1 J-index Algorithm

First, we assume a collection of scientific literatures as a directed graph G = (N; E) in which each node e in N represents an article and each edge (u; v) in E indicates a citation from paper u to paper v. Our goal is to find a metric F(.) such that F(e) represents the academic influence of paper e.

And here we have three assumptions, 1. A paper's academic influence increases as it gains more citations.2. A paper with stronger citations intends to be more influential.3. A paper cited by more innovative papers is more influential.

We assume that the words from a paper are all from K topics, we named the topic from X1 to Xk. And the word of a paper has two sources. First, it creates the word on his own, second, the word was cited from other papers of this certain field. The possibility of this problem satisfy a certain distribution. One word maps one topic. We will get a distribution of topics about one paper. And then get a score of a paper.

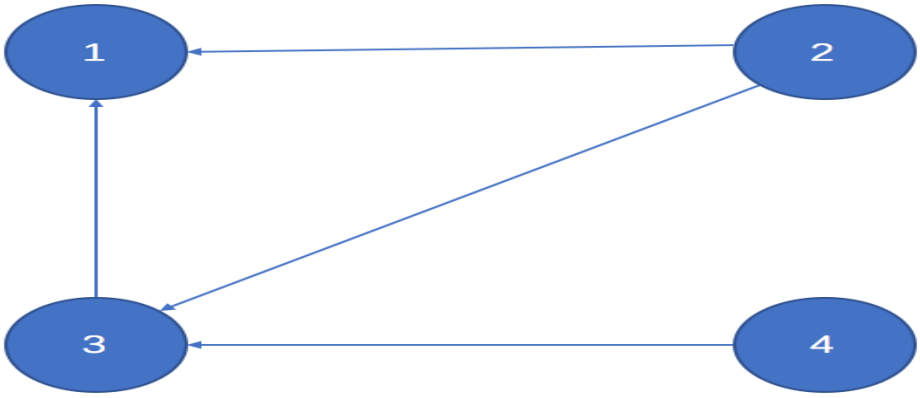
Apart from this, the topics' connections also give us the information about the references among the examples. We can know the relationships among papers through this way directly.

## 2.2 Page Rank Algorithm

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. A PageRank results from a mathematical algorithm based on the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular one. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank

metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

In this project, Page Rank Algorithm is used to calculate the value of papers by the reference relationship. The core point is that a paper cited by more innovative papers is more influential. Just as the second assumption in J-index.




(just as we can see, First we give all the point the same initial value, and then iteration, the edge condition is: (sum of difference)/(sum of sum/2) < 0.001)

2.3 Word Frequency Algorithm

In this model, we assume the words Frequency of one field also has influence in the papers' value. As we can see, there must be some same words among the words from all the papers of this field. And we assume that the words frequency tells us the value of the word to some extent.

We sort all occurrences of the frequency of the occurrences. And then give different grades to the word we sorted. The word appears more, the grade of the word is higher. Thus, we can acquire a distribution of word scores. One word appeared has one score. And then we give the scores of the words back to the paper. After adding all the words of one paper. We can get a paper grade. And this grade tells us the value the papers to some extent.

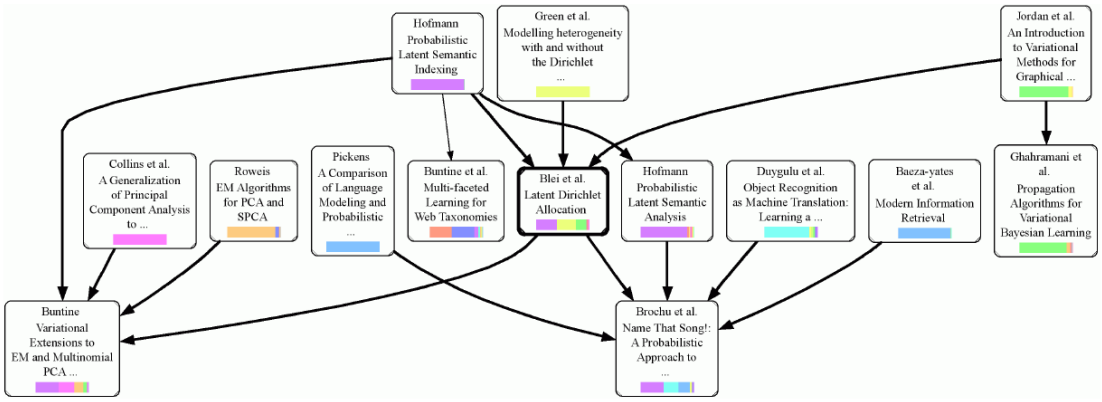
# 3.Graphical Design

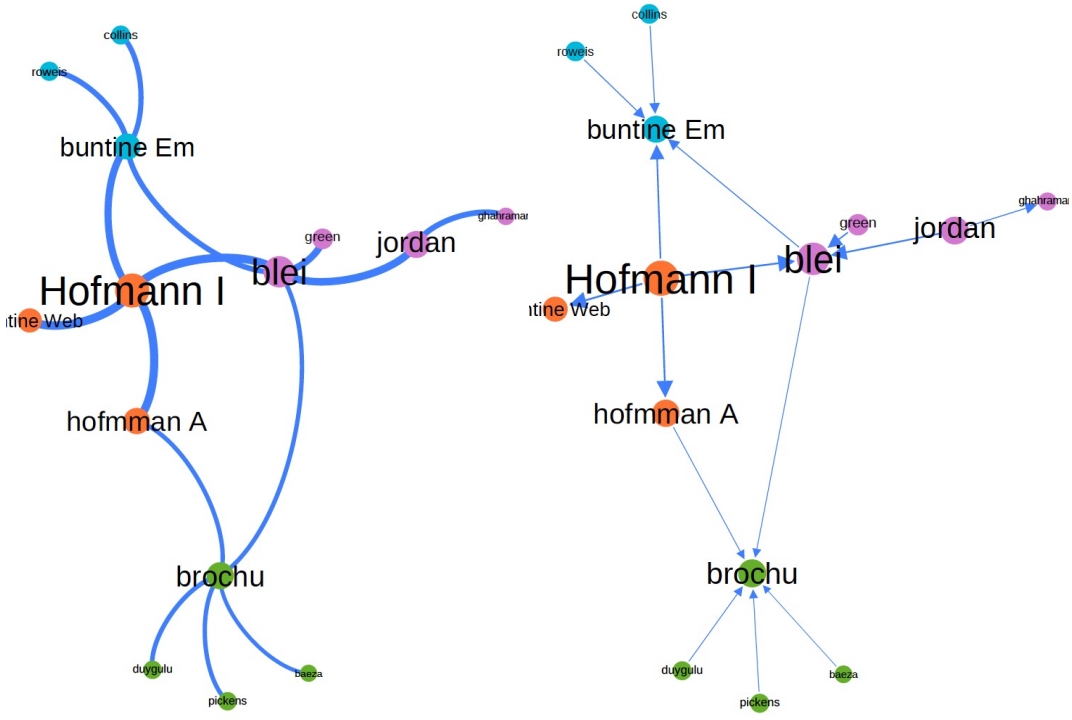
As we can see, graph is more intuitive than grade. So if we want to know the relationship of papers and see the value of the papers directly. Graph is the best choice we can choose.

We draw the result in a graph obeying the following principles:

1. The grade of the paper is combined by the following formula:
   Grade = (J-index Grade)*45% + (Page Rank Grade)*45% + (Word Frequency Grade)%10
2. The weight of the edge is on behalf of the reference strength of two papers.

Here is an example of the total Algorithm:



# 4.Normalization

We must admit that there may be coincidence between two fields. And at the same time, we want to compare two papers from different fields. But we can conclude that through the way we used to weigh the paper. Even if two papers from two fields get the same grade. Their levels may not at the same height. So when we want to compare two or more papers from different fields. We need to normalize them. And I will introduce my normalization way.

As we can see, the best paper in A field may not the same as that one in B field. But we can assume that the top 10% and the last 10% have the similar level. So we assume that the papers grade range in A field is (a,b), and in B field is (c,d). Then we map (a,b) and (c,d) to (x1,x2) together.

In this way, we don't destroy the distribution of any field. But we must admit that if two fields' distributions differ much. We can't use this way to evaluate the value and the connection. Assume a extreme assumption, one field has millions of papers, however, another only dozens of papers. We can't judge the value in this way. The way is reasonable to some extent, but it also has condition to use.