REPORT

Project Title : Predicting Significance of publication using bibliographic text data

Team Members : Text_analysers (201405546, 201507632,201512659)

Introduction:

The impact and significance of a scientific publication is measuredmostly by the number of citations it accumulates over the years.Early prediction of the citation profile of research articles is a significant as well as challenging problem.In this project we tried to predict the significance of a publication using the bibiliographic text data,author information and features derived from citations to predict the long term behavior.

Dataset:

The dataset used by us is DBLP-Citation-network V7 dataset.It has 2,244,021 papers with 4,354,534 citation relationships.The dataset is around two years old.This dataset along with the standard DBLP data also contains the details of which paper refers which paper.For some papers,it also contains abstracts of the research papers.

The dataset is designed in a very easy to use and understand way. It is a very informative dataset but some things which we needed for this project like number of times a paper is cited in another paper, etc. are missing. So, we had to restrict ourselves at some points.

The dataset format is as follows:

The details about a research paper as are described in a block. Each block conatins many rows describing the data.Each row indicates a different attribute of the paper and starts with a different indicator.

- #* - paperTitle
- #@ - Authors
- #t – Year of publication
- #c - Venue of publication
- #index 00- index id of the paper
- #% - the id of references of the paper (there are multiple lines, with each indicating a reference)
- #! - Abstract of the paper

Features Extracted

To do the feature extraction from the data we first converted the data into json format. Each object represented a research paper.

Then, extracted three types of features on the basis of the dataset.

- Author information based features
- Paper information based features
- Citation number based features

Some features were directly computed from data, whereas some required additional computation.

Features Based on Authors Information

The author of a publication plays an important role in its popularity. The following four author-centric features were used for citation prediction:

- **Author h-index (Hindex):** This feature measures average h-index of the authors at the time of publication.
- **Author productivity (ProAuth):** Author productivity refers to the count of his publications. The feature is an average of the productivity of the all the co-authors of a paper.
- **Author diversity (AuthDiv):** Author diversity refers to the diversity in the research fields of author publications. The feature is an average of all the authors taken together.
- **Sociality of author (NOCA):** This feature counts the number of co-authors in all the publications of each author present in the paper.

For all the above the features, we have assumed that all the papers published by the author till date are present in dataset.

Features based Publication properties

The following are the features extracted on the basis of the details about the reasearch papers:

- **Team Size:**The number of authors in a paper
- **Reference Count:**The number of references mentioned in the reference section of a paper.
- **Reference diversity (RDI):** RDI measures the diversity in the fields of the referred papers. A paper citing papers of various fields has a high value of RDI.

Citation number based features

These features helped us to analyse on how the publication numbers have changed over the years after the publication of the paper.

For this we calculated five values which gave us how many papers have referenced the paper after these many years of publication.

We divided the timeline after publication of paper into five parts.

- <=3 years after publication
- 3> and <=6 years after publication
- 6> and <=9 years after publication
- 9> and <=12 years after publication
- >12 years after publication

Citation Profiles

We have assumed that the importance of a paper can change in the following five ways. On this basis only we have divided the papers into 5 possible classes. These classes define how the popularity of paper has changes over the years after it's publication.

The five classes are :

- **PeakInit(PI):** Papers whose citation count peaks within 3 years of publication followed by an exponential decay.
- **PeakLate(PL):** Papers having very few citations at the begining and then a single peak after at least 3 years of the publication ollowed by an exponential decay in citation count.
- **MonDec(MD):** Papers whose citation count peaks in the immediate next year of the publication followed by a monotonic decrease in the number of citations.
- **MonIncr(MI):** Papers having a monotonic increase in the number of citations from the very beginning of the year of publication till the date of observation.
- **Oth(Other):** Paper not belonging to any of the above mentioned categories belong to this category.

Model Construction

We used SVM based classifier to classify the papers into different categories. For training and testing purposes we took a randomly selected sample of 10000 papers from the original dataset of around 3,50,000 papers.The training set was of size 8000 and testing set was of size 2000.The training was done on SVM with Linear kernel and RBF kernel.We also did 5-fold cross validation.

SVM with Linear Kernel

Correctly Classfied : 1983

Misclassified : 16

Classes	**PI**	**PL**	**MD**	**MI**	**Oth**
PI	29	0	0	3	1
PL	0	12	1	0	2
MD	0	4	1547	0	1
MI	1	0	0	258	0
Oth	3	0	0	0	137

SVM with RBF kernel

Correctly Classfied : 1627

Misclassified : 372

Classes	**PI**	**PL**	**MD**	**MI**	**Oth**
PI	1	0	0	0	0
PL	0	0	0	0	0
MD	19	12	1557	209	118
MI	4	0	0	55	7
Oth	1	1	0	1	14