

Oral vs Poster Publication Influence Analysis

Zhengyuan Yang
Department of Computer Science
University of Rochester
zyang39@cs.rochester.edu

Jiebo Luo
Department of Computer Science
University of Rochester
jluo@cs.rochester.edu

ABSTRACT

Many top CS conferences in Artificial Intelligence assign accepted papers into different presentation types, mainly orals and posters. There is a perception that oral papers are better, however some paper with high influence later were also accepted as poster. To end this discussion, in this paper, we proposed a framework to analysis the influence of presentation type and other factors on academic works. To be specific, statistic analysis and significant test are done based on publication's citation count. Further, a better influence representation was proposed based on the ranking methods in [15]. In which influence level was calculated based on the influence of other works and authors in its citation network, instead of simple applying citation count as an influence measure. According our analysis on conference, we show that oral are statistically better than poster in most conferences. Also our ranking method can better represent paper's influence comparing to simply using citation count.

Keywords

Data Mining, Text Analysis, Academic Search

1. INTRODUCTION

Artificial Intelligence has become a hot topic since 1960s. During year hundreds and thousand of great works were generated by talented scholars. These papers lies in an interesting way along with time, topic and presentation type. In this project, we want to reveal some of these interesting patterns with data mining techniques.

In many conferences in AI area, especially in computer vision, accepted paper are assigned into different presentation type. Oral and Poster is a most common way of separating presentation type, though other types include spotlight, poster short introduction, etc. And there is a common perception that oral papers are better than posters. However, it can be noticed that lots of very influential works are accepted as posters at that time. For example:

- A decision-theoretic generalization of on-line learning and an application to boosting.[2]
- Action recognition with improved trajectories.[13]

Further, some conferences claims that all papers are accepted with the exact same criteria, which are later divided into oral and poster based on the estimated interest level and the concern of topic balance. Therefore, we start to concern whether oral are statistically better than poster, or it is just a perception.

To begin with, citation count was applied as the only indication of the influence level for the publication. Statistical analysis and significance test are done this this part, including various pre-processing techniques. Based on that the conclusion is that oral papers are statistically better than poster papers in most of the conferences. Further, a regression model was trained to predict the citation count based on publications attributes.¹ This work is related to citation count prediction[1, 6, 14, 12], as done by all the academic service provider[7, 10].

Though citation count is one of the key value of representing publication's influence, it fails to take citation networks with weight. According to common sense, publication cited by other high influential works are more possible to be better than those who are cited by works with less citation themselves, though the citation count of two works might be the same. Similar idea was also mentioned in [9, 4, 10, 16], though it was done in a different approach. Based on that assumption, we proposed a ranking method to take the citation count as a weighted value and generate the new influence represent, this work is inspired by [15]. According to our experiment on the collected dataset, this new identification can better represent the influence of a academic publication.

The rest of the paper are organized as follows. Section 2 provide an introduction on the preprocessing and statistical analysis methods applied, a short introduction on the collected dataset was also given in this part. Section 3 describes the details of the new influence score generated by the ranking method. Section 4 includes various experiments on our collected dataset. Conclusion was given in section 5.

2. IS "ORAL" BETTER?

In this section, the collected dataset and the analysis based on that was introduced. Problem discussed in this section is whether oral works are better than posters. Some related

¹Haven't completed this part yet.

Table 1: Dataset Introduction

| | year | Oral | Poster | Total | Ratio |
|----------|-------------|------|--------|--------|--------|
| CVPR | 07-16 | 848 | 3451 | 4297 | 1:4.08 |
| ICCV | 07-15 | 204 | 1443 | 1647 | 1:7.07 |
| ECCV | 06-14 | 178 | 1208 | 1384 | 1:6.86 |
| ABOVED 3 | - | 1226 | 6102 | 7328 | - |
| ICPR | 06-10,14 | 948 | 2093 | 3041 | 1:2.21 |
| BMVC | 03-05,08-15 | 303 | 818 | 1121 | 1:2.70 |
| ACCV | 07,10,12 | 96 | 443 | 539 | 1:4.61 |
| ABOVED 3 | - | 1347 | 3354 | 4701 | - |
| TOTAL | - | 2573 | 9456 | 12,029 | - |

topics, e.g. conference quality, influence of best paper, were also discussed.

2.1 Dataset Description

Accepting paper as different presentation type happens most frequently in computer vision conferences, therefore currently the collect data are all from computer vision conferences. The presentation type of the publication (oral/poster) was extracted from the conferences’ web-pages. Cvpapers.com also provide parts of organized presentation type information. As each conference has different way of informing oral/poster, this part of information was retrieved semi-automatically.

After getting the paper name and presentation relation, further information including [name, year, name of conference, author, citation counts][11] was retrieved with Microsoft academic API[10]. Currently the dataset including more than 12,000 papers from recent years’ CVPR, ICCV, ECCV, ICPR, etc. Detailed number in each presentation was given in table 1.

2.2 Citation Count based Influence Analysis

In this part, the applied pre processing and statistical analysis method was introduced, which works on the data got from computer vision conferences and citation count was used as the identification of the publication’s influence. As oral and poster session is also a part of a conference, we are discuss generally on conference quality measurement, which could be used to analyze the influence of different presentation type.

2.2.1 h5-index

h5-index is one of the most frequently used index to judge the quality of a conference. h5-index is the h-index[3] for articles published in a certain conference during the last 5 years. h-index is the number of papers h which has at least h citation. Table 2 gives an example of the h5-index given by Google Scholar in the computer vision and pattern recognition track. We also start with applying h5-index to show the influence of each conferences and their oral/poster session.

H-index gain a good success in evaluating researcher’s scientific output, with its balanced consider of total number of paper, total number of citation, citation per paper and ”significant papers” with citation over $num_{threshold}$ or top k of that person’s work. And because of that, it can avoid disadvantages in each individual criteria above may have, for example failing to focus on the significant papers, totally ignore the less influential works, or problems in picking the threshold ideally.

Table 2: h5-index given by Google Scholar

| Rank in cs.CV | Publication | h5-index |
|---------------|---------------------|----------|
| 1 | CVPR | 140 |
| 2 | TPAMI | 114 |
| 3 | ICCV | 92 |
| 4 | TIP | 92 |
| 5 | arXiv (cs.CV) | 87 |
| 6 | ECCV | 76 |
| 7 | Pattern Recognition | 67 |
| 8 | IJCV | 65 |
| 11 | BMVC | 41 |
| 20+ | ACCV | 26 |

h5-index, during 2011-2015.

However, the direct apply of h5-index on computer vision areas have two potential problems. First, Total number of papers should not be considered as a factor for quality in conference analysis. Unlike severing as the productivity criteria for researcher, the total number of papers have less relation with the quality of the conference. Some top conference like KDD accept small number of works. Further, some conferences hold every two years(e.g. ICCV, ECCV) and it is unfair to compare the h5-index directly with other every-year based conferences. Second, h-index was first brought out in physics and also been proved to be useful in many other areas like biology. However, as the great development in AI area, the paper might have a different distribution than h-index assumed. One obvious phenomenon is that computer scientists, especially in AI, seem to have a higher h-index.

2.2.2 Distribution and Statistics

As h5-index failed to capture all the information in a conference, further analysis are done with distribution and statistics.

Citation per paper Citation per paper is the most simple size independent measurement of the dataset. The disadvantage of using average to measure researchers, that it fails to penalize the low productivity, is no longer a problem in conference quality measurement. There are mainly two problems. First is that there are frequently some ”outlier” in citation count, which have extremely high citation (e.g more than 5,000). And it will greatly elevate the citation per paper. Though it could be solved by normalization, we believe it will change the original distribution and make no sense, thus we just leave it there. Second problem is that recent papers tend to receive less citation. What make the problem even worse is that some conference didn’t hold on the same year and becomes impossible to compare by average without pre processing. For example, citation per paper of ECCV2014 is higher than ICCV2015, however it can’t be told whether it is the affect of time, or the quality of the conference. Therefore, a model was proposed to reduce the influence of time. It output a weight vector generated based on the data we got and make different year’s work comparable. Results on ICCV2015 before and after weighting is shown in Fig 1.

Distribution Though citation per paper provide a general knowledge on the conferences’ quality, it fails to represent the essential information of conferences’ inner distribu-

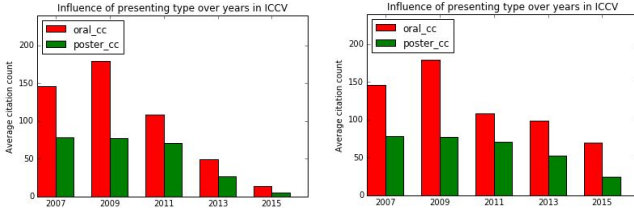


Figure 1: Pre processing to reduce the influence of time. Shown on ICCV 2015 data. Left is the original value and right is the weighted one.

tion. Whether a session or conferences have higher average because of more highly cited outliers? Or because there are less works with low citation? These are all important information to gain, in order to understand the quality of the session.

Box plot was generated for distribution analysis. It provides the 25% value, 75% value and the median. Rest two values are selected as $Q_1 - 1.5 \times IQR$, $Q_3 + 1.5 \times IQR$. Based on box plot a rough idea about the citation distribution can be obtained. Further, the histogram of citation count with different width of bins was generated to have a closer study of the distribution. The histogram without binning was fitted with Pareto Distribution, which has probability density function as shown in 1.

$$p(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \text{ for } x \geq x_m \quad (1)$$

x_m was selected as 1 and α was changed to fit the data. A smaller α will lead to a distribution with smaller value near zero while having a longer tail when x becomes larger. Pareto Distribution was selected because of it fits the power-law like distribution we observed. In addition, we could focus on the difference in paper with low citation (e.g. less than 3) and very high citation with modifying the α parameter.

2.2.3 Distance Measure and Significant Test

After getting an understanding about the session and conference, following two questions are brought up and solved in this part. First, whether orals are statistically better than posters in all and in each conferences? Second, if it is true that orals are better, which conference did the best in picking out the good paper for oral presentation?

To solve the first problem, t-test was done between orals and posters in each conference following 2, and results are compared with threshold on different confidence level.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_\Delta} \quad s_\Delta = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (2)$$

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)} \quad (3)$$

The knowledge of orals are statistically better than posters are gained through the t-test. Based on that, we calculated the distance between the probability histogram of orals and posters to answer which conference can better pick the good papers. Based on the conclusion that orals are better than posters, the assumption that smaller difference between histogram indicates oral are less better than poster in that

conference, therefore implying a lower ability of picking out good works. Cosine distance, supremum distance and KL-divergence (equ:4) was calculated.

$$D_{KL}(P||Q) = \sum P(i) \log \frac{P(i)}{Q(i)} \quad (4)$$

2.3 Citation Network Based Influence Value

2.3.1 General Idea

In section 2, analysis for publication influence based on citation count was discussed. Though citation count is a good represent for the publication's influence, it consider all the citation to be equally weighted. According to common sense, a citation from another famous paper, or a influential scholar should be more "valuable" than a normal citation. Based on this idea, we proposed a new represent for publication influence with the following assumptions.

1. A citation² is important if many important scholar cited it and it was cited by other important publications.
2. A scholar is important if he or she has a high h-index. Meanwhile his or her publication have high citation count and have received important citation.
3. A publication is important if it has a high citation count. Meanwhile it has been written by an important scholar and have received important citation.

A ranking algorithm was proposed to solve this problem, this work is inspired by [15, 5]. And the citation relation information was extracted from Google Scholar.

2.3.2 Ranking Algorithm

The key value in this algorithm is the importance value of citation, scholar and publication, represented as P_C, P_S, P_P . P_S is the importance value for scholars, the length is equal to the number of scholar collected in our data. P_P is the importance value for publication, the length is equal to the number of publications in dataset. P_C is the importance value for publication's citation, its length is equal to the number of total publications and representing the weighted sum of all citation transaction's contribution. These importance vectors following the rules.

$$P_P = M_{PC} \cdot P_C \quad P_S = M_{SP} \cdot P_P$$

$$P_C = M_{CS} \cdot P_S \quad P_S = M_{CS}^T \cdot P_C$$

$$P_P = M_{SP}^T \cdot P_S \quad P_C = M_{PC}^T \cdot P_P$$

M_{PC} is the matrix between publication and citation. For a publication's citation, if it was cited by publication A, the corresponding value in the matrix will be set to the value of A's citation count, which represent publication A's initial influence. If it was not cited by publication B, the corresponding matrix value will be left zero. Similarly for user and citation matrix M_{CS} , the value in the matrix will be set to the author's h-index. The reason of setting to h-index and citation count related value instead of binary value will be

²the citation here is the total information of a publication, about it was cited by which publication and scholar. Instead on a one-to-one transaction.

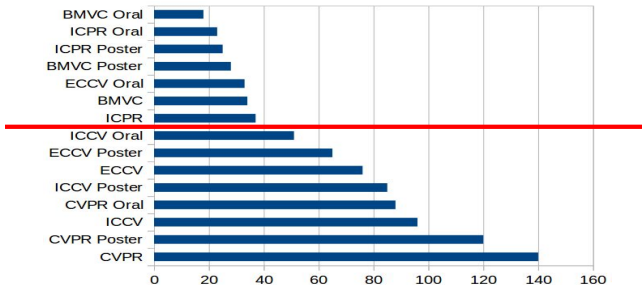


Figure 2: h5-index on each conferences and their sessions.

discussed in next part. Matrix M_{SP} represent publication was written by which scholars.

By iteratively update P_C, P_S, P_P . P_S , the algorithm will converge on a value for P_C , which include the contribution of all other papers which cited it and is the new influence value we proposed.

2.3.3 Converge and Discuss on Initializing

The above algorithm iteratively find citation importance value vector P_C , which is the eigen value for $M \cdot M^T$, where $M = M_{PC} \cdot M_{SP} \cdot M_{CS}$. According to [8], the convergence is guaranteed.

The initialized value of P_C, P_S, P_P . P_S is not important, as long as P_C is not orthogonal. As we want to make use of the prior knowledge on author and publications' influence, which are commonly represented by h-index and citation count, we wish we could find a method to start iteration based on that prior knowledge. Because initializing P_S, P_P won't work mathematically, we embed the prior knowledge to M_{CS}, M_{PC} . Therefore, the importance value of the publications and scholars are not limited to the data we got, but are extracted from the whole citation network and fine-tuned on our specific area (computer vision), which make the final citation influence score more reliable.

3. EXPERIMENTS

Based on the discussion in section 2, experiments are done on the computer vision dataset we collect. Dataset description can be found in section 2.1.

3.1 h5-index

As mentioned, h5-index is a good identification of conferences or session's quality. From Fig 2 shows the good quality of the top three conference in computer vision: CVPR, ICCV and ECCV.

Besides, as h5-index is greatly influenced by the total number of publications, the result is misleading in certain aspects. We could notice the posters have slightly higher h5-index than orals, and CVPR seems much better than ICCV and ECCV. However, it might be mainly caused by the difference in size. For example, CVPR accepted 4297 papers during 2011-2015, which is roughly the same as the sum amount of ECCV, ICCV. This is mainly because CVPR hold and year while the rest two hold every two year. Thus if CVPR 2012 and CVPR 2014 was used to calculate h5-index, the new value will be 96 which is much closer to ECCV's h5-index. Similarly, the h5-index generated by CVPR 11,13,15 is 119. The size problem also happens on oral and poster comparison.

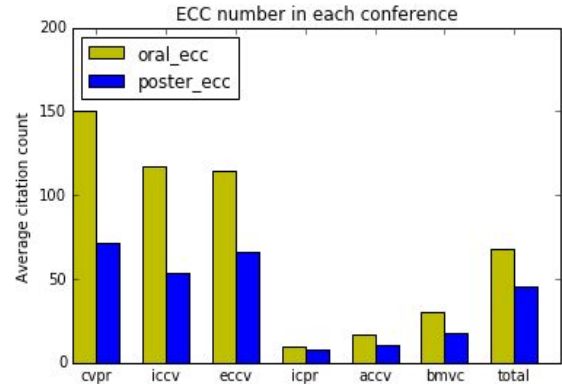


Figure 3: Citation per paper in each conference and session.

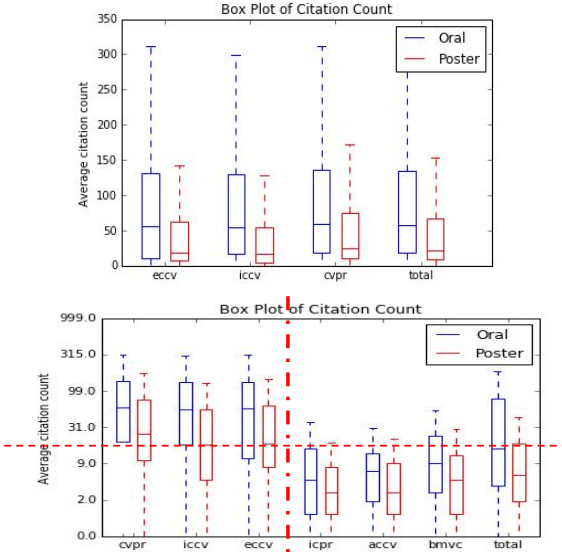


Figure 4: Box Plot of the distribution. Upper one is generated on the whole collected data, and the lower one is based on the top three conferences in CV.

3.2 Citation Per Paper

The average citation count in each conference and session was given in Fig 3. Though a significant outstanding is observed in every conference, we much discuss too much on citation per paper as it fails to display the reason of difference and whether the difference is statistically significant enough.

3.3 Distribution and Fitting

Box Plot Box plot was given in Fig 4. Comparing to oral with posters, a significant difference could be noticed that orals are better. And from the box plot the difference is mainly because posters have much more works with low citation count. This will be further analyzed with histogram.

Pareto Distribution Fitting As shown in Fig 5, Pareto distribution can fit the citation count histogram pretty well. As the distribution fitting poster have a large α value than orals, it also prove the guess that one big difference between oral and posters is there are far less publication with low citation.

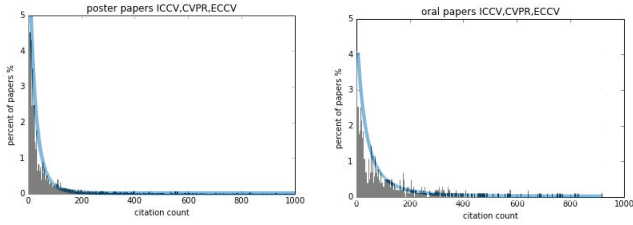


Figure 5: Histogram on citation count on top 3 conferences without binning. And the Pareto Distribution used to fit them. Left is the poster data while the right one is oral.

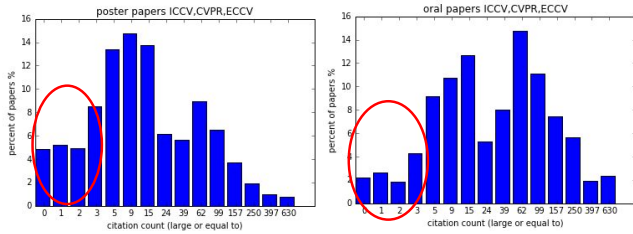


Figure 6: Binned histogram with logarithmic coordinates. Notice the difference in red circle.

Binned Histogram To further prove the guess, the binned histogram was generated with logarithmic coordinates (Fig 6). It further prove the difference in lower part.

Difference on Influential paper Another finding is inspired by the long tail of both oral and poster distribution, that we analyze the difference between publication with very high citation (e.g top 3 %). According to the result in Fig 7, though there is a significant difference between the top 10% of two presentation types, the gap becoming smaller and smaller for the top 1,3,5%. It proves that great works can be collected by both oral and poster.

3.4 Distance Measure and Significant Test

3.4.1 Significant Test

Though according to the analysis in section 3.3, oral seems to be better than poster, further experiments have to be done to ensure whether the difference are significant enough. T-test was done in all the conferences as show in Fig 8, from which we could tell that orals are statistically better than posters in almost all the computer vision conferences.

3.4.2 Distance Measure

In this part, analysis are done on figuring out which con-

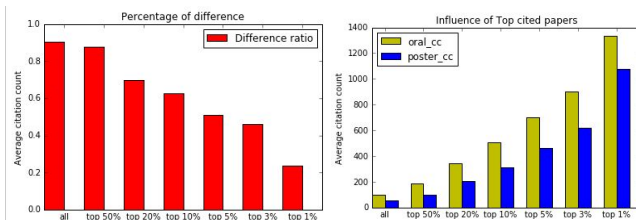


Figure 7: The average citation count and the difference between oral and poster for the top k % publications.

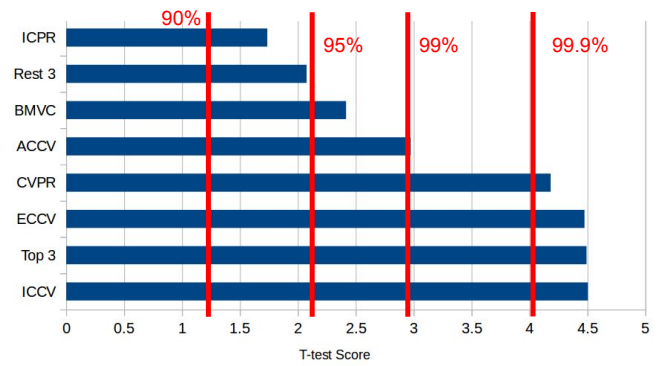


Figure 8: T-test on each conference with visualization of threshold on different confidence value. Degree of freedom various from 14 to 18, threshold was plot with d.f. set to 15.

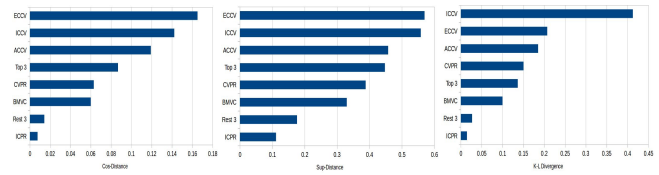


Figure 9: The probability histogram distance between the distribution of orals and posters. The plot from left to right is cosine-distance, supremum-distance and KL-Divergence.

ference can better pick the good papers as oral presentations. Cosine distance, supremum distance and KL-Divergence were selected as the similarity measurement to show how different orals are comparing to posters. With the prior knowledge gained in section 3.4.1, orals are statistically better than posters, we make assumption that the distance represent how much orals are better, thus reflecting conference’s ability in picking orals.

From the results it shows that top conferences can better pick orals. Though the following two factors should also be noticed. First, the significance is influenced by the amount of publications chosen as oral. If the percentage goes high, it becomes more and more difficult for the orals to be significant. That partly explains why orals in CVPR are not that outstanding comparing to posters in CVPR. As the oral:poster ratio which is 1:4.08 is relatively low, comparing to 7.07 in ICCV and 6.86 in ECCV. Also, some conference will pick orals concerning the balance and diversity in presentation, which is another factor of oral picking besides the publication’s quality.

4. CONCLUSIONS

In this paper, the analysis on conferences’ presentation type is carefully studied. In our collected computer vision publication dataset, following four conclusions are drawn. (a). Three top conferences in computer vision(CVPR, ICCV, ECCV) are statistically better than the rest. Among which CVPR is slightly better than rest two. (b). Oral papers are statistically better than Posters in computer vision track. Mainly because there are less none-influential works in orals. (c). Both poster and oral could have influential works. (d). Top conferences seems to have better taste on picking oral presentations. In addition, we proposed a novel importance

value for representing the publication's influence based on the citation network. With the consideration of scholar and citing paper's influence, the new importance measure can better represent the quality of the publication comparing to citation count.

5. REFERENCES

- [1] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.
- [2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.
- [3] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005.
- [4] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.
- [5] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, and J. Han. Reinforced similarity integration in image-rich information networks. *IEEE transactions on knowledge and data engineering*, 25(2):448–460, 2013.
- [6] M. Khabsa and C. L. Giles. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949, 2014.
- [7] H. Li, I. Councill, W.-C. Lee, and C. L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, pages 883–884. ACM, 2006.
- [8] T. H. Michael. Scientific computing: an introductory survey, 2002.
- [9] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.
- [10] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246. ACM, 2015.
- [11] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.
- [12] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics, 2006.
- [13] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.
- [14] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1247–1252. ACM, 2011.
- [15] Z. Yin, L. Cao, J. Han, J. Luo, and T. S. Huang. Diversified trajectory pattern ranking in geo-tagged social media. In *SDM*, pages 980–991. SIAM, 2011.
- [16] X. Zhu, P. Turney, D. Lemire, and A. Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427, 2015.