Analysing Trends in Computer Science Research

A Preliminary Study Using The Microsoft Academic Graph

Suhendry Effendy
School of Computing
National University of Singapore
effendy@comp.nus.edu.sg

Roland H.C. Yap
School of Computing
National University of Singapore
ryap@comp.nus.edu.sg

ABSTRACT

Research in Computer Science (CS) evolves rapidly in a dynamic fashion. New research area may emerge and attract researchers, while older areas may have lesser interest from researchers. Studying how trends evolve in CS can be interesting from several dimensions. Furthermore, it can be used to craft research agendas. In this paper, we present trend analysis on research area in CS. We also look at citation trend analysis. Our analysis is performed using the Microsoft Academic Graph dataset. We propose the *FoS score* to measure the level of interest in any particular research area or topic. We apply the FoS score to investigate general publication trends, citation trends, evolution of research areas, and relation between research areas in CS.

Keywords

trend analysis; computer science; bibliographic databases

1. INTRODUCTION

Research in Computer Science (CS) evolves rapidly in a dynamic fashion. New research areas can emerge while older areas become less popular but yet such older areas can also get a resurgence. Studying how trends evolve in CS can be interesting from several dimensions: (i) it is itself worthy of study in its own right as a kind of meta-research of CS itself; (ii) it shows how different areas in CS are inter-related; and (iii) it can inform about new trends and popularity of certain research areas. Furthermore, understanding of trends can itself be used to set research agendas.

One difficulty with studying trends is that enough data is required, as such a good source of data is needed. Bibliographic data with authors, title and conference/journal information is not so difficult to obtain in CS, e.g. DBLP [1]. However, other kinds of data like keywords, citations, topics can be much harder to obtain in any *comprehensive fashion*. Furthermore, determining the topics of a paper is itself a research problem [4]. Recently, *Microsoft Academic* provides a rich source of data making it easier to get at such data.

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053064

In this paper, we propose that rich data sources of academic papers such as Microsoft Academic can be used to study research trends. In particular, Microsoft Academic has an analysis which classifies the research topics of papers into fields of study (FoS). We show that analysing the FoS can show interesting trends of research areas as a whole or from the perspective of citations in Computer Science. We also investigate trends in particular conferences, evolution of research areas and inter-relationships between research areas in CS. We also propose a scoring measure, the FoS score, to measure the contribution of a paper in terms of its topics. While this paper is only a preliminary study using the MAG dataset (a snapshot of the data in Microsot Academic), we believe those general aspects of the trends will also apply to the real Microsoft Academic data which is continually updated.

1.1 Related Work

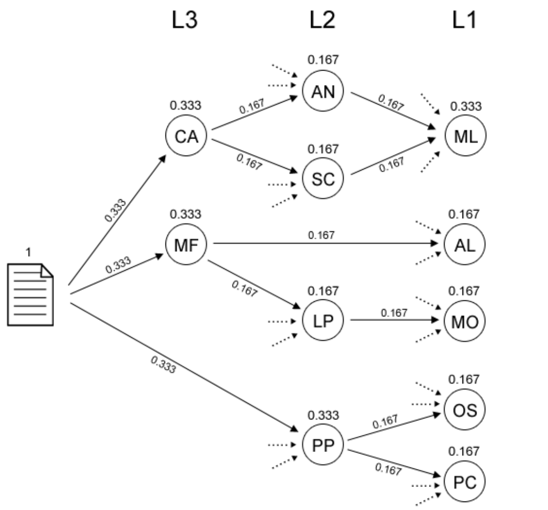
There are not many works which study systematically trends in CS. Hoonlor et al. study trends in CS especially its relation with research funding [5] using papers from ACM and IEEE with research areas based on ACM and IEEE classifications. We employ a dataset from *Microsoft Academic* with a much larger set of papers focusing on conferences (1,716,211 papers in 1,279 conferences). More importantly, the areas (field of study) do not come from pre-classification but through semantic analysis. Other related works studying CS conferences are on the relation between conferences [2] and conference categorization [4].

2. MICROSOFT ACADEMIC GRAPH

In this paper, we employ a snapshot of bibliographic data provided by Microsoft Academic[1], *Microsoft Academic Graph* (MAG). MAG was released by Microsoft Research in 2015 [8] originally as part of the WSDM challenge [11]. MAG contains various information about publication (e.g., papers, authors, venues, keywords, citations) which are obtained via crawling. Since the source data is crawled, MAG may contain errors and be incomplete. For example, there may be errors in the publication venue, author information, etc. In this paper, we assume that while they may be noise in the data, it is not significant to the trends studied.

MAG contains papers across many disciplines. In this paper, we want to focus our study on CS papers. MAG maps

[1] `http://academic.research.microsoft.com`



CA: *cluster analysis* (L3); MF: *maximum flow problem* (L3); AN: *artificial neural network* (L2); SC: *statistical classification* (L2); LP: *linear programming* (L2); PP: *parallel processing* (L2); ML: *machine learning* (L1). AL: *algorithm* (L1); MO: *mathematical optimization* (L1); OS: *operating system* (L1); PC: *parallel computing* (L1).

Figure 1: Example of FoS paper score from one paper with three keywords, i.e. on clustering algorithms, maximum flow, and parallel processing.

each paper into its *field of study* (FoS)[2], thus, we can roughly discern the topic of each paper without needing to analyse the abstract of the paper or the paper content itself. Each paper may have multiple FoS, however, MAG does not measure the importance of each FoS in the paper. For example, papers published in a biology journal may use or borrow some technique from the computer science field; thus, the papers may have the FoS given as *biology* and *computer science*. On the other hand, papers published in a computer science conference could also use techniques from another field, e.g. mathematics, and similarly have both FoS. In both examples, MAG does not distinguish the importance of each FoS. Intuitively, we do not consider papers in the first example as CS papers as their main focus is on biology. However, papers in the second example should be considered as CS papers.

One way to find CS papers is by considering only papers published in CS publication venues. Unfortunately, there is no simple way to isolate publication venues specific for CS research. In this paper, we do not consider a white-list of such conferences. Rather, we use the observation that research in CS is mostly published in conferences [7, 9]. Although there are CS papers in journals, important results are often published first in conferences. Hence, in this paper, we obtain our set of CS papers by considering only papers which have their FoS given as *computer science* and are published in conferences.

In MAG, each paper has a set of keywords, however, there is no unique ID for each keyword and it only serves to bridge the paper to FoS. Thus, we consider each paper as having a set of FoS in MAG where each FoS represents a research area or a topic. There are 4 levels of FoS, i.e. L0 to L3, with L0 being the most general FoS, e.g., *computer science*, and L3 being the most specific, e.g., *k-nearest neighbors algorithm*. In general, the hierarchy of the FoS is in the form of directed acyclic graph instead of a tree, i.e. a FoS may have more than one parent FoS. For example, *k-nearest neighbors algorithm* (L3) FoS belongs to *statistical classification* (L2), *artificial neural network* (L2), and *machine learning* (L1). It turns out that keywords for a paper in MAG need not be mapped to the lowest FoS level (L3 FoS), e.g. a paper may be tagged with *data mining* (L1 FoS).

2.1 FoS Score

In this study, we leverage on the FoS provided in MAG for each paper. One simple way to do trend analysis is simply by *counting*, e.g., count the number of papers, number of citations, etc. However, each paper in MAG may have multiple FoS, which may bias the results in favour of papers with many FoS. We propose instead to use another measure, the *FoS score* (or simply *score*) for analysing trends in FoS. In particular, we define the *FoS paper score*, $\Gamma(f)$, to measure the effect of a paper. Similarly, the *FoS citation score*, $\Upsilon(f)$, to measure the contribution of incoming citations for each FoS. We start by defining the FoS paper score and later adapt it for the FoS citation score.

For each paper, we distribute the scores to the affected FoS such that each paper contributes exactly 1 to the total score. This score is then propagated to the higher level FoS. The FoS paper score is defined as follows. Let P denote the set of all papers and $F_B(p)$ is the set of FoS directly mapped from paper $p \in P$. In MAG, the F_B does not necessarily contain L3 FoS only. The score for each FoS f obtained directly from paper p is defined as:

$$\Lambda(p,f) = \begin{cases} 1/|F_B(p)| & \text{if } f \in F_B(p) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In Equation 1, the score from a paper p is distributed equally to all $f \in F_B(p)$. The FoS paper score for FoS f from a paper p is defined recursively as:

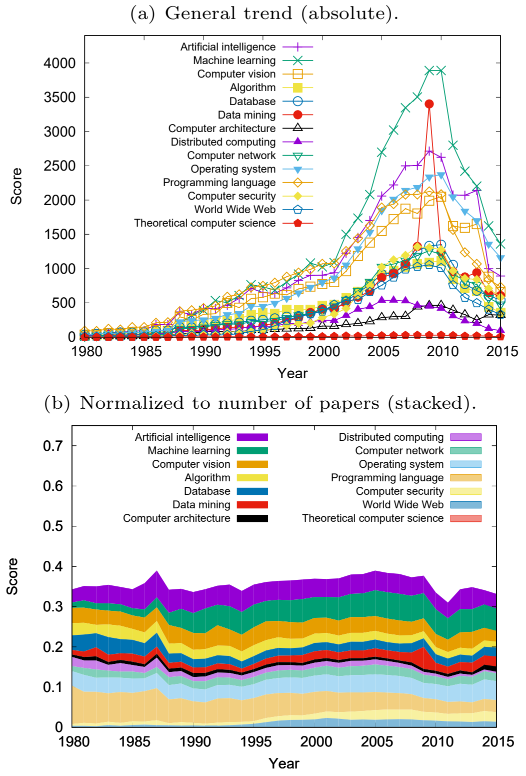
$$\Gamma_p(f) = \Lambda(p,f) + \sum_{g \in \pi(f)} \frac{\Gamma(g)}{|\rho(g)|} \tag{2}$$

where $\pi(x)$ is the set of all direct child FoS of x, and $\rho(x)$ is the set of all direct parent FoS of x. The first term of Equation 2 is obtained directly from paper p, while the second term is the propagated score from f's children. Notice that the score for an FoS g is propagated equally to its parents. Finally, the *FoS paper score* for a particular FoS can be defined as:

$$\Gamma(f) = \sum_{p \in P} \Lambda(p,f) + \sum_{g \in \pi(f)} \frac{\Gamma(g)}{|\rho(g)|} \tag{3}$$

In this paper, as we are only studying the CS research area, in the experiments we only consider FoS which are direct or indirect children of FoS *computer science*. For this reason, $\rho(x)$ contains only all parent FoS of x which are direct or indirect children of FoS *computer science*. Figure 1 shows an example of FoS paper score from one paper, $\Gamma_p(f)$. The FoS hierarchy in Figure 1 is extracted from MAG where all affected FoS and their parents are shown for the selected paper, while sibling and other child FoS are not shown as they are not relevant in the example (represented by the

[2] FoS are also used in *Microsoft Academic Service* [10].


(a) General trend (absolute).

(b) Normalized to number of papers (stacked).

Figure 2: Trend of FoS paper scores ($\Gamma(f)$) for selected L1 FoS in MAG.

dotted lines in Figure 1). For example, the direct parent FoS of *maximum flow* (i.e. ρ(*maximum flow*)) is *algorithm* and *linear programming*, while *machine learning* has multiple children where two of them are *artificial neural network* and *statistical classification*. In Figure 1, the paper has three keywords which are mapped directly into three FoS: *cluster analysis*, *maximum flow problem*, and *parallel processing*, each with score of $\Lambda(p, f) = 0.333$. These scores are then propagated to their respective parent FoS (see Figure 1). After the propagation, the scores on the L1 FoS are: *machine learning* (0.333), *algorithm* (0.167), *mathematical optimization* (0.167), *operating system* (0.167), and *parallel computing* (0.167).

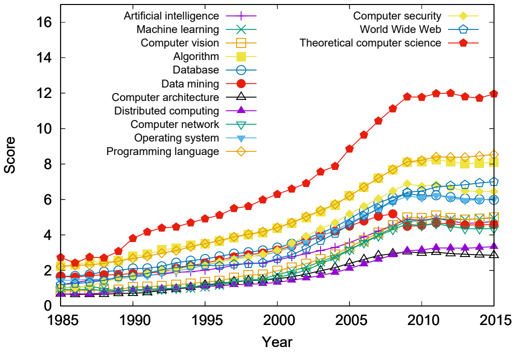
To analyse the incoming citation trend, we adapt Equation 3 to consider the citation data. Citations are "directed references" in the form of (q, p) which means paper q cites paper p. Namely, p receives an incoming citation from q. Thus, the *FoS citation score* is defined as:

$$\Upsilon(f) = \sum_{(\cdot,p)\in R} \Lambda(p, f) + \sum_{g\in\pi(f)} \frac{\Upsilon(g)}{|\rho(g)|} \tag{4}$$

where R is the set of all citations and $(\cdot, p)$ indicates an incoming citation for paper p.

3. PRELIMINARY RESULTS

In this section, we discuss some preliminary findings to illustrate what can be obtained by trend analysis in MAG. We caution that MAG is only a snapshot and while it has a broad selection of papers the data may neither be complete nor clean. In our analysis, we only consider papers in MAG with FoS *computer science* and published in conferences. The MAG dataset consists of information on 1 354 603 papers, 1 324 591 authors, and 1277 conferences. This is smaller compared to DBLP with ∼1.8M papers in proceedings but is only a snapshot. In addition to the bibliographic data, MAG also labelled each paper with its FoS. There are also 2 455 223 incoming citations to the papers.



Data point on year Y comprises all citations/papers published no later than year Y.

Figure 3: Trend of ratio between FoS citation scores ($\Upsilon(f)$) and FoS paper scores ($\Gamma(f)$) on selected L1 FoS in MAG.

3.1 General Trend

To illustrate basic statistics and trends which can be seen from FoS on MAG, we performed basic analysis on various research areas in CS as represented by their FoS. One seemingly easy way to do this is simply by calculating the score for each FoS in each year using the method in Sec. 2.1. Figure 2(a) shows the result on selected L1 FoS in MAG. These FoS are selected to represent a broad subset of research areas in CS. One should exercise care when interpreting this result. It seems generally there is an increasing trend on the presented FoS; however, this might be due to the increasing number of papers and conferences. We also remark that the drop from 2011 is simply because there is fewer data in that period in MAG.

An alternative visualization is shown in Figure 2(b) which is normalized by the number of papers in each year. The peak in machine learning in 2(b) becomes much less dominant compared with 2(a). However, the clear trend in the growth of research in *machine learning* can be seen, and also for *computer security*. We can see areas like *programming language* which have become less "popular" over time. On the other hand, there are topics which are not highly popular but are quite stable, e.g., *computer architecture*. These trends would seem to consistent with expectations.

We can see that the distribution of papers in 1980 is more skewed compared to later years, in part, this is due to growth across other areas. Note that in MAG, there are 34 L1 FoS in *computer science*. We only show 14 selected FoS in Figure 2 to make the visualization easier to see.

3.2 Citation Trend

Another way to study trend is by examining the citation trend, in particular, the incoming citation trend. We computed the score for each FoS using Equation 4. We analyse the trend using the ratio between $\Upsilon(f)$ (FoS citation score)

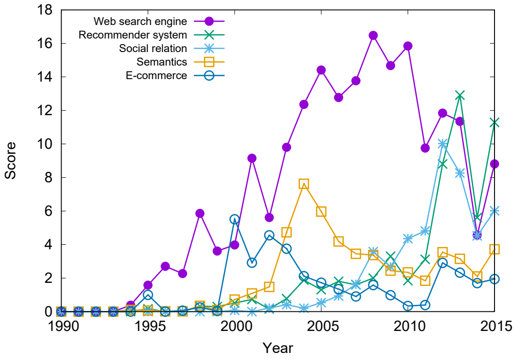


Figure 4: Trend of selected L2 FoS in WWW.

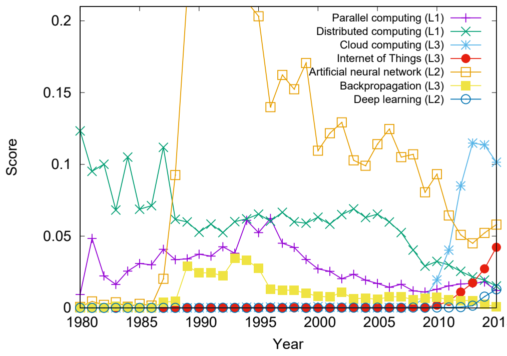


Figure 5: Trend of selected FoS in MAG.

and $\Gamma(f)$ (FoS paper score). Specifically, for each particular year Y, we calculate the ratio between $\Upsilon(f)$ from all citations from papers published in year Y or before, and $\Gamma(f)$ from all papers published in year Y or before; in other words, it is the incoming citations score normalized by citable papers score for each particular FoS. This measure is intended to show which research area is more influential in term of citations. $\Upsilon(f)$ only considers the incoming citations regardless of the number of citable papers. An FoS may have high $\Upsilon(f)$ simply because there are many papers with such FoS (reflected in a high $\Gamma(f)$); thus, $\Upsilon(f)$ should be normalized by $\Gamma(f)$ if we want to analyse the influence of f in term of citation.

Figure 3 shows the result. As we can see, *theoretical computer science* has the highest ratio; even though there are considerably fewer papers in *theoretical computer science* (Figure 2(a)). Essentially in MAG, there are many incoming citations to papers with a *theoretical computer science* FoS. This is not unexpected as "theory" often becomes a foundation for other research. We further analyse the data on the *theoretical computer science* FoS. We found more than 85% citations to papers with a *theoretical computer science* FoS come from papers not labelled as *theoretical computer science* FoS. This is consistent with the reasoning above.

We see that although there are many papers which have a *machine learning* FoS (Figure 2(a)), however, it seems the citations are not as high as other FoS. We found that less than 55% of citations to papers with *machine learning* FoS come from papers without the *machine learning* FoS. There are some FoS, such as *computer architecture* and *distributed computing* receive lesser citations on average compared with other FoS. However, CS evolves, and the *machine learning* FoS has grown in citations since the 1990s. We believe this is a reasonable way of measuring impact or popularity. Some areas are naturally higher impact while others may be more specialized.

We also observed some potential anomalies in MAG, A paper can cite another unpublished (at the time) paper. We highlight that this may or may not be an error, for example, the citation could be to a paper archive, e.g. arXiv. However, there are only ∼0.2% of such citations involving ∼0.6% papers, thus are likely not significant to the overall trends.

3.3 Trends in Particular Conferences

Trend analysis for research area in a particular conference is arguably clearer and easier to do as the number of conferences is fixed (i.e. a single conference). The number of accepted papers in each year may affect the trend, but the trend is not due to the changing number of conferences which can occur in the general case.

In this section, we present an analysis on WWW (the World Wide Web Conference). Figure 4 shows the trend for the top-5 L2 FoS in WWW, i.e. FoS with the highest score. As we can see, *web search engine* is a dominant topic in WWW until 2012 when *recommender system* and *social relation* emerge in popularity. Research in *e-commerce* hits its peak in the year 2000, We remark that this is around the same time frame when the dot-com bubble bust. Using this kind of simple analysis, we suggest that one can learn what topics are popular in a certain conference and how they evolve through time.

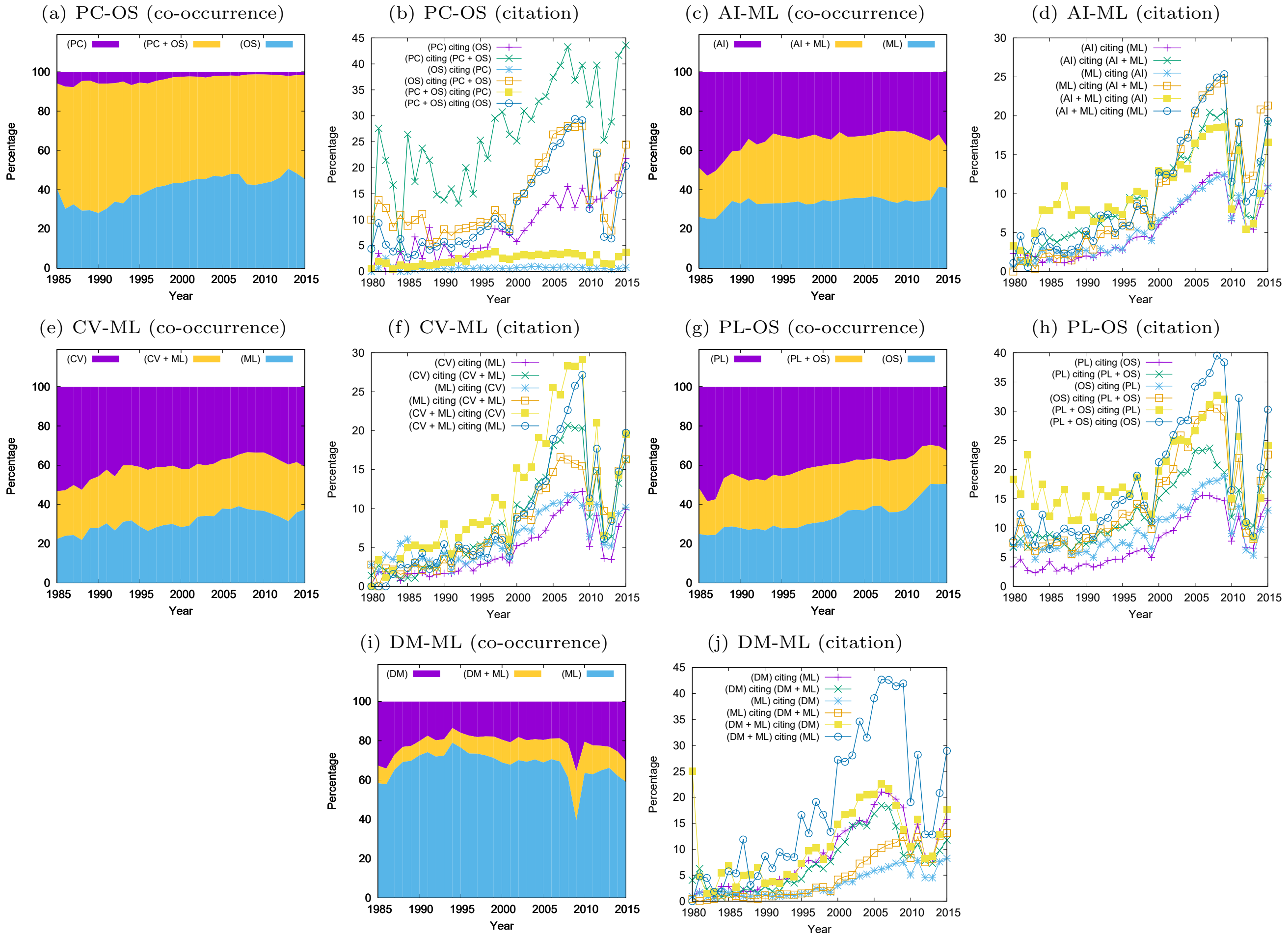
3.4 Evolution of Research Areas

Naturally, research areas in CS evolve over time. New areas may emerge and attract many researchers, while older areas may have lesser interest. In this section, we present an analysis on some selected FoS. Figure 5 shows the trend of FoS paper score with the FoS at various levels. First, observe that *distributed computing* has a stable trend up to 2006, and then decreases; around a similar timeframe, related areas such as *cloud computing* and *internet of things* emerge and become popular. We note that both *cloud computing* and *internet of things* are not children of *distributed computing* in the MAG FoS hierarchy.

We remark that trend analysis relies on having enough data. Recently, *deep learning* is popular but this is not shown in MAG as there is insufficient recent data in the snapshot. We see a decreasing trend of *artificial neural network* from 1993 onwards and a rebound around the same time as the appearance of *deep learning*. However, without enough data, it is less clear whether the rebound is caused by *deep learning* although we might expect that to be a possible reason. Note that both *artificial neural network* and *deep learning* are L2 FoS in MAG, thus, they do not have a parent-child relationship.

3.5 Relation between Research Areas

Different research areas may also interact with each other. We can see this through a simple kind of interaction such as co-appearance in a paper, or a more complex interaction such as citation flow. Table 1 presents the top-5 L1 FoS co-occurrences in MAG and their number of papers. As we can see, there are many papers with both *parallel computing* and *operating system* FoS; comprising ∼15% of the dataset.



PC: *parallel computing*; OS: *operating system*; PL: *programming language*; AI: *artificial intelligence*; ML: *machine learning*; CV: *computer vision*; DM: *data mining*.

Figure 6: Relation between research areas.

Table 1: Top-5 L1 FoS Co-occurrence

# papers	FoS$_1$	FoS$_2$
228K	*parallel computing*	*operating system*
205K	*programming language*	*algorithm*
197K	*artificial intelligence*	*machine learning*
183K	*computer vision*	*machine learning*
167K	*programming language*	*operating system*

This is followed by *programming language* - *algorithm*, and *artificial intelligence* - *machine learning*. These combinations seem very natural. For example, intuitively there is a likely relationship between the *parallel computing* FoS and *operating system* FoS. However, this table does not show the direction of the relationship — does *parallel computing* influence *operating system*, or is it the other way around?

Figure 6 presents a more detailed visualisation of the relationship giving the direction for some selected pair of FoS. In Figure 6(a) and 6(b), (PC + OS) refers to papers which have FoS on both *parallel computing* and *operating system*, (PC) refers to papers which have FoS on *parallel computing* but not *operating system*, and (OS) refers to papers which have FoS on *operating system* but not *parallel computing*. Other abbreviations and symbols are given in the caption of Figure 6.

In Figure 6(a) we can see that most papers in *parallel computing* also have *operating system* FoS – labelled as (PC + OS) in the figure; only a small fraction of papers in *parallel computing* do not have *operating system* FoS – labelled as (PC). In Figure 6(b) we can see that most papers in (PC) cite papers which have both (PC + OS) FoS. Moreover, there is only a small fraction of papers in (OS) which cite papers in (PC). This result suggests that research in *parallel computing* is more influenced by *operating system* than the other way around.

Figure 6(c-h) on *artificial intelligence* - *machine learning*, *computer vision* - *machine learning*, and *programming language* - *operating system* are examples where both FoS seem to be more equal in term of influence on each other. Figure 6(i-j) presents the relation of *data mining* (DM) - *machine learning* (ML). Although that *data mining* and *machine learning* do not appear in top-5 co-occurrences in Table. 1, we present their comparison as these two topics are often perceived as being closely related. Notice that there are more papers in *machine learning* compared to papers in *data mining* as shown in Figure 6(i). In Figure 6(j) we can observe that there are more citations to papers in *machine learning* from papers in *data mining* than the other way around. This may suggest that *machine learning* is more influential to *data mining* than *data mining* to *machine learning*. However, the influence may not be as strong as *operating system* to *parallel computing* as in Figure 6(a-b).

4. CONCLUSION

In this paper, we present several trend analyses in Computer Science using the Microsoft Academic Graph (MAG) data, in particular, using the FoS and citation data. We propose a way to do the analysis by exploiting the hierarchical field of study (FoS) provided by MAG for each paper. Although we investigated MAG, having some kind of FoS hierarchy is quite natural. As each paper may have multiple FoS and each Fos may have multiple parent FoS, we propose to do trend analysis using the FoS score. This allows analysis on any FoS of interest (by propagating the score to the higher level FoS); it also prevents bias on papers which have many FoS.

From the experiments, we see that many intuitive trends can be readily seen by straightforward analysis in MAG. This has potential since this is a dataset which is constantly being expanded and improved upon. We also present and discuss results on citation trends. This analysis may also be useful for other kinds of analysis of scholarly data, for example, perhaps one application is towards conference rating prediction [6, 3].

5. ACKNOWLEDGMENTS

This work has been supported by Microsoft Research Asia. We also acknowledge the assistance of the Microsoft Academic team to help us better understand the MAG dataset.

6. REFERENCES

[1] DBLP. The dblp computer science bibliography. `http://dblp.uni-trier.de/`.

[2] S. Effendy, I. Jahja, and R. H. C. Yap. Relatedness measures between conferences in computer science - a preliminary study based on DBLP. In *WWW Workshop on Big Scholarly Data*, 2014.

[3] S. Effendy and R. H. C. Yap. Investigations on rating computer sciences conferences: An experiment with the microsoft academic graph dataset. In *WWW Workshop on Big Scholarly Data*, 2016.

[4] S. Effendy and R. H. C. Yap. The problem of categorizing conferences in computer science. In *Theory and Practice of Digital Library (TPDL)*, 2016.

[5] A. Hoonlor, B. K. Szymanski, and M. J. Zaki. Trends in computer science research. *Communications of the ACM*, 56(10):74–83, 2013.

[6] I. Jahja, S. Effendy, and R. H. C. Yap. Experiments on rating conferences with CORE and DBLP. *D-Lib Magazine*, 20(11/12), 2014.

[7] J. A. Konstan and J. W. Davidson. Should conferences meet journals and where?: A proposal for 'PACM'. *Communication of the ACM*, 58(9):5–5, 2015.

[8] Microsoft. Announcing the microsoft academic graph: Let the research begin! `https://www.microsoft.com/en-us/research/blog/announcing-the-microsoft-academic-graph-let-the-research-begin/`, 2015.

[9] D. S. Rosenblum. The pros and cons of the 'PACM' proposal: counterpoint. *Communications of the ACM*, 58(9):44–45, 2015.

[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. WWW Companion Volume, pages 243–246. ACM, 2015.

[11] A. D. Wade, K. Wang, Y. Sun, and A. Gulli. WSDM cup 2016: Entity ranking challenge. In *Web Search and Data Mining (WSDM)*, 2016.