A Tale of Two Communities: An Analysis of Three Decades of Case-Based Reasoning Research

Barry Smyth(✉)

Insight Centre for Data Analytics, School of Computer Science,
University College Dublin, Dublin, Ireland
barry.smyth@ucd.ie

Abstract. We analyse three decades of case-based reasoning (CBR) research to better understand the health of CBR and its relationship to adjacent research fields. We identify two largely separate CBR communities, one based on the research published at mainstream CBR venues (ICCBR, ECCBR etc.), the other encompassing CBR work with no direct connection to these venues. We analyse their scale, impact, and focus, and the potential to bring them closer together in the future.

1 Introduction

This year the case-based reasoning community returns to Otzenhausen, Germany, 26 years after the first European Workshop on Case-Based Reasoning [14], which many regard as *the* formative event in the history of CBR. The 1993 Otzenhausen meeting led to a long-running series of workshops and conferences, as EWCBR became ECCBR, and later merged with ICCBR. Returning to Otzenhausen is a natural time for community *reflection* and the purpose of this paper is to support this by analysing more than 600,000 articles, including CBR papers, their referenced and citing papers, and other (non-CBR) papers by CBR authors. It echoes, and expands upon, similar analyses carried out in the past [5–7], while at the same time introducing new ideas about how we might evaluate the state of CBR today.

During ICCBR community meetings, one frequent topic for discussion concerns the existence of another CBR community without a close connection to mainstream CBR venues. This discussion often arises in the context of how we might increase the size of ICCBR, attract additional submissions, and otherwise further accelerate the development of CBR. We investigate whether such a community exists – spoiler, it does! – and we compare and contrast the scale of activity across both communities: their output and impact; the topics they

Supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

K. Bach and C. Marling (Eds.): ICCBR 2019, LNAI 11680, pp. 343–357, 2019.
https://doi.org/10.1007/978-3-030-29249-2_23

emphasise; the most influential ideas that have emerged etc.[1]. Our aim is to better understand the similarities and differences between both communities and to identify opportunities to bring them closer together in the future.

In the next section we will describe our main publication dataset and how we distinguish between the two CBR communities mentioned above. Subsequent sections examine publication output, community dynamics, and citation impact. We also describe the results of a topic modelling and citation analysis in order to identify the principal research themes, and the most influential papers, which have emerged from the last three decades of research, and more recently.

2 Datasets and Communities

This work begins with a dataset provided by Semantic Scholar (SS[2]), which provides publication meta-data and citation data for more than 46 m publications, primarily from the fields of computer science and health science.

2.1 The CBR Dataset

We select a *CBR dataset* of 675,118 papers by 1,042,490 unique authors from the following subsets of SS data:

- V_p, the set of *venue* papers; papers published at I/ECCBR and EWCBR.
- S_p, the set of papers returned in an SS *search* for CBR papers[3].
- C_p, the full set of CBR papers ($C_p = V_P \cup S_p$).
- L_p, the set of *linked* papers that cite, or are cited by, papers in C_p.
- R_p, the set of *related* (non-CBR) papers by authors of papers in C_p.

There are corresponding sets $(V_a, S_a, C_a, L_a, R_a)$ for the *authors* of these papers. As summarised in Fig. 1(a), there are 8,223 unique CBR papers in C_p, 66,941 linked papers, and 632,770 papers (606,165 + 2,012 + 6,211 + 18,382) authored by CBR authors. The corresponding data for authors is in Fig. 1(b).

[1] We will avoid the temptation to *name-check* individual researchers, on the grounds that such rankings can end up as distractions to the central argument.

[2] SS is an open, research-article search engine; see https://www.semanticscholar.org.

[3] We identified candidate papers based on a set of *strong* (e.g. case-based reasoning, derivational analogy), *moderate* (e.g. case adaptation, case based), and *weak* (CBR, case retrieval, case learning) search terms, and a scoring metric to identify CBR papers with a high degree of accuracy. Due to space restrictions it is not possible to provide a complete account of the terms and weightings used. The process involved considerable trial and error and validation tests were performed to ensure good precision and recall during the final dataset preparation.

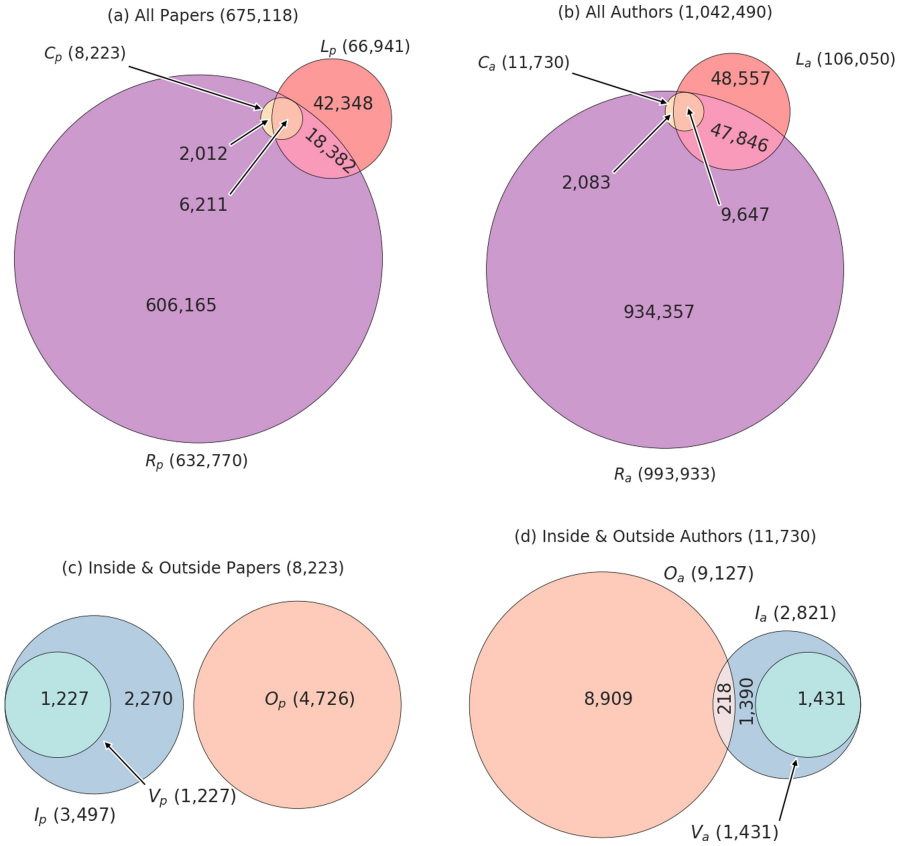


Fig. 1. Venn diagrams of the various datasets and subsets of data used in this study.

2.2 A Tale of Two Communities

A central idea in this work is that the CBR field is shared by two largely separate communities. We refer to CBR researchers connected to the *mainstream venues*[4] as the *inside community*, and to CBR researchers without a direct connection to the mainstream venues as the *outside community*.

More precisely, a CBR paper, p, is in the set of *inside papers*, I_p, if and only if p is co-authored by a *venue* author. And a CBR author, u, is in the set of *inside authors*, I_a, if and only if u is an author of an inside paper; see Eqs. 1 and 2. Notice, that $V_a \subset I_a$; an inside author does not have to be a venue author, but they must *co-author* with a venue author. Thus, inside authors are *connected* to the mainstream venues by venue authors, but they do not necessarily need to publish in the mainstream venues themselves.

[4] We use the term *'mainstream'* to refer to ICCBR/ECCBR/EWCBR, but only as a convenience, and without attempting to impugn the many other research venues where CBR papers appear.

$$I_p = \{p \epsilon C_p \mid \exists u \epsilon Authors(p) \wedge u \epsilon V_a\} \quad (1)$$

$$I_a = \{u \epsilon C_a \mid \exists p \epsilon I_p \wedge u \epsilon Authors(p)\} \quad (2)$$

Conversely, a CBR paper, p, is an *outside paper* if it is not in the set I_p, and u is an *outside author* if u is an author of an outside paper; see Eqs. 3 and 4.

$$O_p = C_p \setminus I_p \quad (3)$$

$$O_a = \{u \epsilon C_a \mid \exists p \epsilon O_p \wedge u \epsilon Authors(p)\} \quad (4)$$

Figure 1(c, d) summarises the number of papers and authors in these inside and outside sets, and the relationships between them. Notice in Fig. 1(c) how the inside papers are a superset of the venue papers ($V_p \subset I_p$). Notice too that the inside and outside papers are mutually exclusive ($I_p \cap O_p = \phi$), but the inside and outside *author* sets are not ($I_a \cap O_a \neq \phi$). In Fig. 1(d) there are 218 authors who are both inside and outside authors.

Each one of these 218 authors is an author of an inside paper but they are not venue authors – they have *co-authored* with a venue author, but only outside the core venues – and each is also a co-author of an outside paper. For example, an author might have been a co-author of a non-venue, inside paper, as a PhD student, then went on to bring their CBR expertise to another group in the outside community, where they also published, becoming an outside author. They are interesting authors because they represent a point of contact between both communities, and may play an important role in creating more contact between these communities in the future. But they are also rare, emphasising the conspicuous lack of connection between both communities.

The scale of the two communities is also worth noting. At the start of this study it was not clear whether the outside community would prove to be more than a limited body of CBR work. Instead we find a significant body of CBR research that is even larger than that of the inside community. The outside community spans a similar period of time but has >35% more papers, and >2x the number of unique authors, when compared to the inside community.

2.3 Inside/Outside Venues

Why are there two, mostly separate, communities? To explore this further, Fig. 2 shows the number of CBR papers at the top-25 most frequently targeted inside and outside venues; for reasons of clarity the mainstream venues have been omitted, as they tend to dominate, making other venues more difficult to compare. We can see that the most popular (non-mainstream) venue for CBR research is *Expert Systems with Applications*, with just over 150 CBR papers during the last 30 years, the vast majority of which (≈80%) have come from the outside community. In contrast, *FLAIRS*, *IJCAI*, and *AAAI* are more frequent targets for the inside community. These are the types of venues – AI, ML, expert systems etc. – that one might expect for CBR publications.

More revealing is a similar plot in Fig. 3, but this time focusing on the *non-CBR* papers published by the inside and outside authors. Now there is a clearer

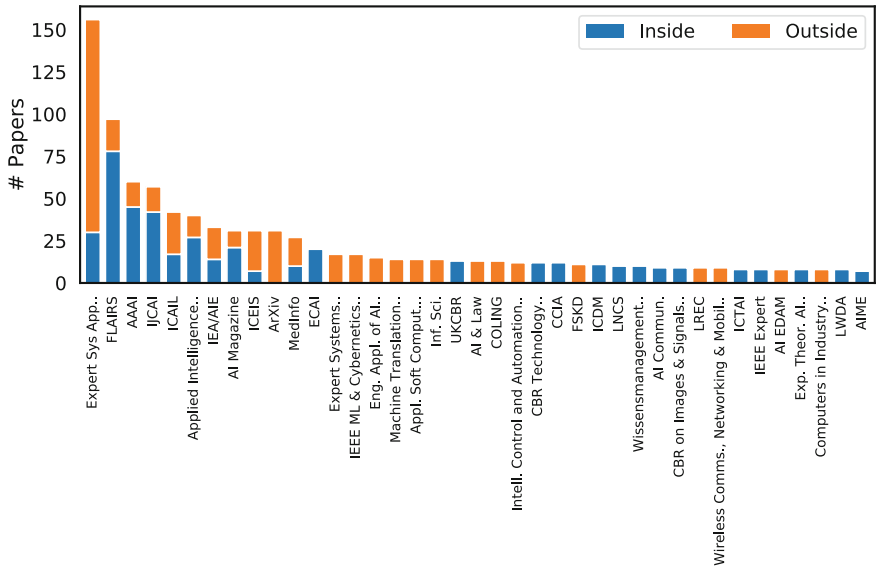


Fig. 2. Top publication venues (excluding I/ECCBR/EWCBR) for inside and outside papers.

division between the top inside and outside venues. The former, as was the case for CBR output, targets mostly AI/ML related venues (*AI Magazine*, *ECAI*, *RecSys* etc.), but the latter targets *PloS One*, *Physics Review*, the *Journal of Biochemistry*, and *Applied Materials* etc. This suggests that while inside authors are mostly AI/ML researchers, outside authors are much more likely to be biologists, physicists, material scientists, and chemists.

The inside community is a community of AI/ML researchers with a focus on CBR, whereas outside researchers come from many different areas, scientific and commercial, using CBR as a technology to solve challenging problems in their home domains. To put this another way, the inside community is *about* a related set of topics (AI, ML, CBR), whereas the outside community is *about* many different topics. As such, we might expect the former to be more coherent and less fragmented than the latter. If so, then there should be a stronger *community effect* for the inside community when it comes to citation impact.

3 Publications, Authors, and Impact

Next we examine the output and impact of these communities, by looking at the volume of publications per annum, the number of active authors, and the citation impact that their work is having.

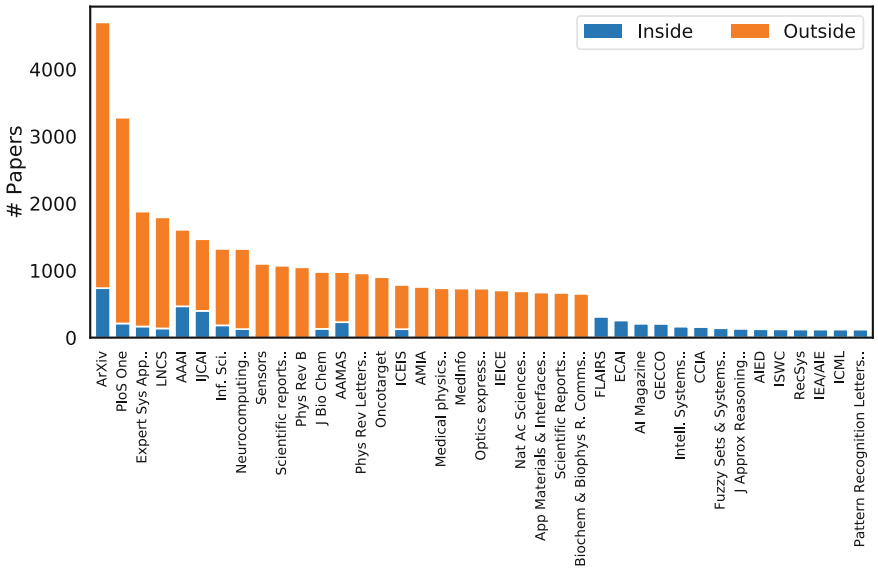


Fig. 3. Top publication venues for non-CBR papers produced by inside and outside authors.

3.1 Publication Output

Figure 4(a) shows the cumulative number of CBR publications produced; for reference, we also show the cumulative output of the mainstream venues. There has been a steady output from each community but since 2009 the total number of CBR papers produced by the outside community has surpassed that of the inside community, a trend that continues to this day.

Approximately 50% of the inside and outside output is made up of application papers; see Fig. 4(b)[5]. CBR has always been an application-oriented field and this is in contrast with a much lower, but growing, fraction of application papers among the non-CBR papers in our dataset, also shown.

3.2 New, Returning, and Churning Authors

While both communities are broadly similar in terms of their publication output, differences begin to emerge when we look at their respective author-bases. Figure 5 shows: (a) the cumulative number of active authors per year; (b) the fraction of new authors per year; (c) the fraction of returning authors; and (d) the fraction of churning/lost authors.

Compared to the inside community, the outside community is characterised by higher levels of new authors and lower levels of returning authors, suggesting

[5] We determine application papers based on the presence of keywords such as 'application', 'domain', 'deploy', for example, in the title or abstract.

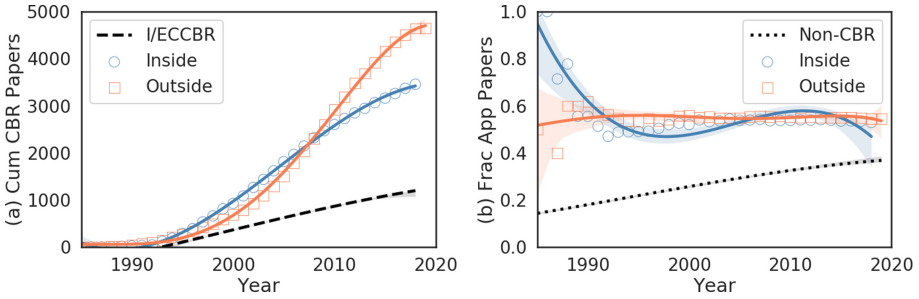


Fig. 4. The (a) cumulative number of CBR papers published per year and (b) the fraction of application-oriented papers per year.

that many outside authors are engaged in CBR research for a shorter period of time. In fact, on average inside authors publish CBR papers over a 5-year period, compared a 3-year period for outside authors; just over 20% of inside authors remain CBR-active for more than 5 years, compared to <10% for outside authors. This is consistent with the idea that the inside community is focused on advancing the fundamentals of CBR – with its researchers engaged for the long-term – while the outside community is a community of practice, with less long-term investment in CBR by its researchers.

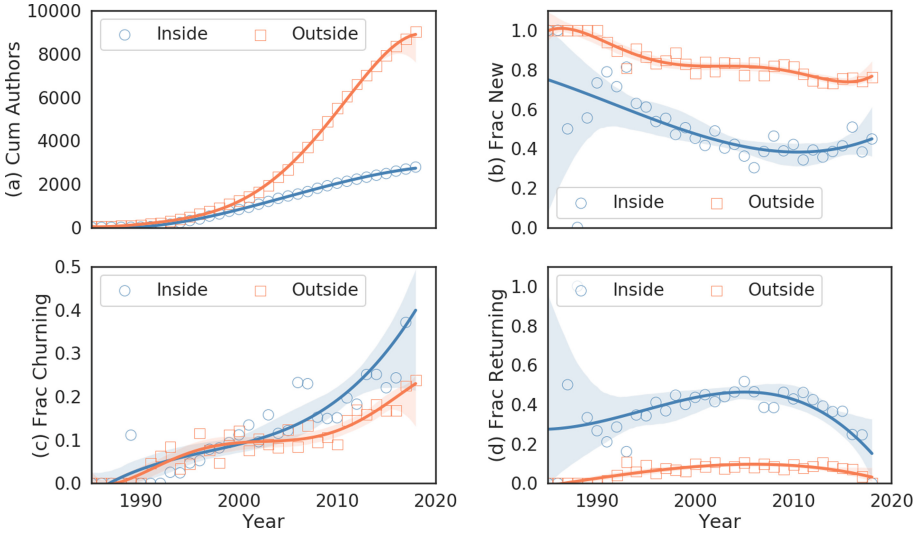
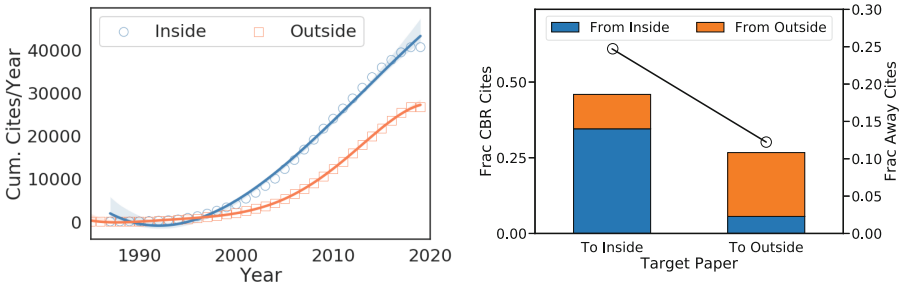


Fig. 5. The (a) cumulative number of publishing CBR authors per year; (b) the fraction of new authors per year; (c) the fraction of churning/lost authors per year; (d) the fraction of returning authors per year.

3.3 Citations and Impact

Another difference between the communities is revealed when we consider citation impact. Figure 6(a) shows a significant citation benefit for inside papers compared to outside papers. From an early point in the development of CBR, inside papers have tended to attract more citations than the outside papers.

Figure 6(b) shows that ≈50% of the citations to CBR papers by the inside community come from other CBR papers, compared to just over 25% for the outside community. Once again, this difference is consistent with the notion that CBR papers by the inside community are more likely to make a central CBR contribution, attracting CBR cites, whereas CBR papers by the outside community are more likely to use CBR in the service of some other task.



(a) Cumulative citations per year. (b) Fraction of cites from CBR papers.

Fig. 6. A citation analysis summary for inside and outside communities.

Figure 6(b), also shows the extent to which one community cites their own work versus the work of the other. We refer to inside papers citing inside papers, and outside papers citing outside papers, as *home cites*. Conversely, inside papers citing outside papers, and outside papers citing inside papers, are *away cites*. Figure 6(b) shows that the inside community benefits from a much higher proportion of away cites (≈25%) than the outside community (≈12%). In other words, outside papers are more likely to cite inside papers than the other way around. All other things being equal, this may suggest a *discoverability* issue for the outside community, which contributes to its lower citation impact; fostering greater links between the community may help to address this.

Figure 7 looks at a number of summary impact metrics, and inside papers continue to benefit. They attract more cites per paper. The time to the first citation is shorter (they are more discoverable). Their citation half-life[6] is longer, and the number of years to peak-cites[7] is greater. Inside papers enjoy a more immediate, significant and sustained impact, compared with outside papers, indicating, as predicted, the inside community benefits from a stronger *community effect*.

[6] The number of years it takes to accumulate 50% of their cites.
[7] The time it takes for the paper to have its best citation year.

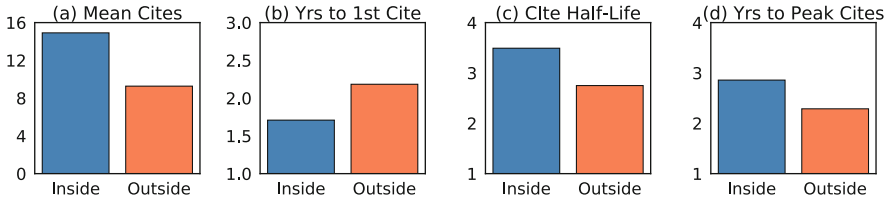


Fig. 7. Summary impact metrics for inside/outside papers.

4 The Evolution of Research Topics

To explore the themes from the last three decades of research we construct topic models for the inside and outside papers by transforming titles and abstracts into term-based representations, using tokenisation, lemmatisation, and stemming. Latent Dirichlet Allocation (LDA) [2] is applied to the resulting *document-term* matrix to produce a *document-topic* matrix (encoding the probability distribution of the topics per document) and a *topic-term* matrix (encoding the probability distributions of the terms per topic). We cluster the papers based on their dominant topics, and use t-SNE (t-distributed stochastic neighbour embedding, [10]) to produce the 2D topic maps shown in Figs. 8 and 9.

In these visualisations each paper is represented by a disc, with papers from the same topic grouped together by t-SNE, and coloured similarly. Distance denotes similarity, the radius of each disc is proportional to the number of citations attracted by the paper, and the opacity of the disc is proportional to the recency of the paper (more recent papers are *more* opaque). Finally, each topic is labeled using the top terms from the LDA probability distributions.

Fig. 8. The inside topics discovered from the inside papers. (Color figure online)

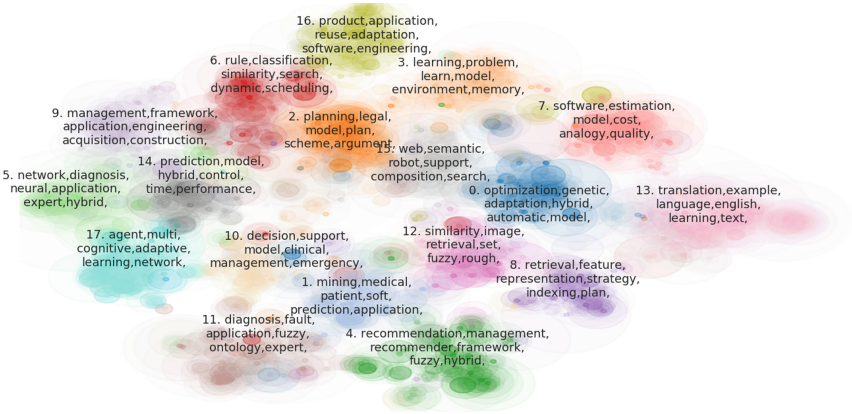


Fig. 9. The outside topics discovered from the outside papers. (Color figure online)

Although there is no direct mapping between inside and outside topics – *inside topic i* is not related to *outside topic i* – there are clearly similarities between the areas explored by both communities, reflecting common themes within CBR research (similarity and retrieval, learning and adaptation, prediction classification, recommendation etc.) But there are differences too. Planning and strategy games (*inside topic 1*), learning and analogy (*inside topic 3*), and maintenance and competence (*inside topic 9*) are important themes within the inside community, but they are less evident among the outside topics. Conversely, the outside topics exhibit a greater emphasis on certain application themes – medical data-mining (*outside topic 1*), software engineering and estimation (*outside topic 7*), and (example-based) translation (*outside topic 13*) – which are less well represented by the inside papers.

Figure 10(a–f) summarises aspects of each of these topics; remember inside topic i has no relationship to outside topic i. The inside/outside topics are similar in terms of their fractions of papers and application papers. Outside topics tend to peak sooner (≈8.5 years) and more recently (≈2013) versus ≈11.5 years and 2007 for inside papers, respectively. The citation benefit for inside papers persists across topics too: inside topics enjoy more cites per paper, and a higher (topic-based) h-index in almost all cases.

There is obviously more that could be explored with respect to the evolution of CBR research topics. It would be interesting, for example, to pay more attention to recent, emerging topics, by building our topic models over a subset of recent papers, or to try and predict future topics. Alternative approaches to clustering papers could also be considered, for example by using co-citation, rather than term-overlap, as a measure of inter-paper similarity. For now we will leave these as open ideas for future work.

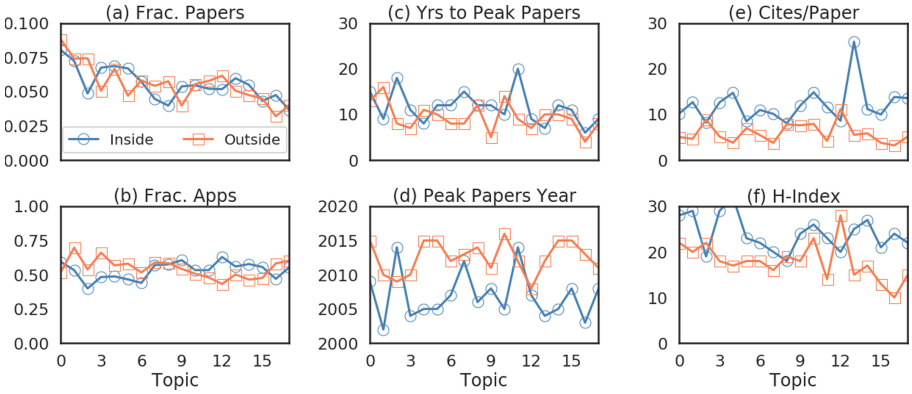


Fig. 10. Summary of various aspects of the inside and outside topics.

5 Influential Papers

In any research area there are some papers that stand out as particularly important to the evolution of the field. Sometimes they are breakthrough papers that introduce new or improved techniques, or even new research directions. Sometimes they are survey papers that bring together a body of research, perhaps reframing it, or integrating it with relevant ideas from other fields. Here we seek to identify the most influential CBR papers, over the past three decades of CBR research, as well as those that have emerged more recently.

5.1 Link Analysis and Influence Metrics

Important papers tend to stand out as being among the most cited works in a field, but citation count alone is not always sufficient to identify the *most* influential articles. In recent years, *link analysis* techniques have been used to evaluate the importance of nodes in a graph, based on various features of network topology. For example, algorithms such as *PageRank* [3] and *HITS* [8] consider a node to be important if it is connected to other important nodes. It is common to use these ideas to reveal influential papers in a *citation graph*, where the nodes are papers and the edges are the citation links between them [4].

We build a single citation graph based on all of the CBR papers (from both communities) and implement three different scoring metrics: (1) the number of cites that the paper has attracted; (2) the *PageRank* score of paper; and (3) the *HITS authority score*. Each metric generates a single score for a paper, which we convert into a rank, and then we calculate the sum of these ranks to generate an overall ranking; using ranks is a simple but effective way to combine these scores in a scale-free manner.

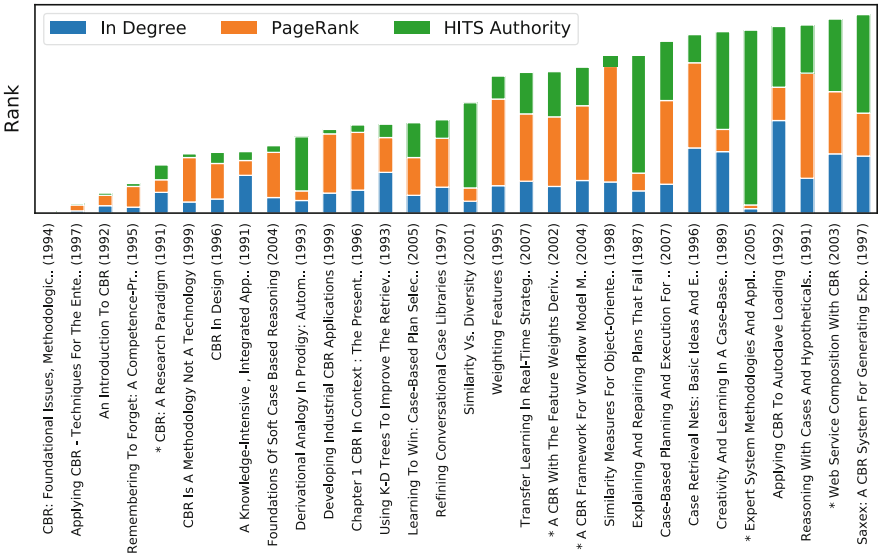


Fig. 11. Top-25 most influential CBR papers (inside & outside communities).

5.2 Seminal CBR Papers

Figure 11 shows the top-30 papers based on this overall ranking. Each bar corresponds to a single paper and shows its ranking across all 3 metrics. Outside papers are indicated with an asterisk prefixing their title on the x-axis. Over one-third of the publications are survey papers or introductory books, including 4 of the top-5 [1,9,11–13]. Only 5 (16%) of the most influential papers come from the outside community and most of the top-30 come from the very early years of CBR research; the mean publication year is 1997. This is not so surprising, as many of these papers established the foundations of the field, and their impact has been building over a long period of time, but it begs the question as to where future influential papers are likely to come from.

5.3 Emerging Influencers

To shed some light on this, Fig. 12 presents a similar set of ranking results, but focusing on the most influential CBR papers just from the last decade (2009–2019). There are far fewer survey papers – notwithstanding that the top ranked paper is a survey of CBR in health sciences – and there is an abundance of outside papers; 15 of the top 30 are outside papers. Perhaps the outside community will prove to be more influential over the coming years.

There is also evidence of a number of increasingly important and novel application domains among these more recent papers. For example, 7 of the papers focus on healthcare and clinical applications (from classical diagnosis and classification to duty rostering), 5 of the papers focus on financial applications

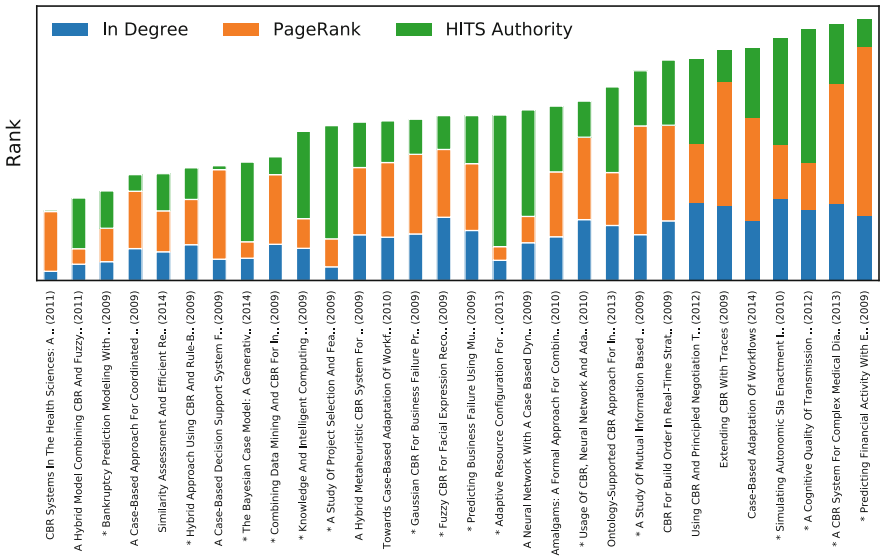


Fig. 12. Top-25 most influential recent (>2008) CBR papers (inside & outside communities).

(including bankruptcy and business closure prediction), 3 consider the application of CBR to workflows, while others explore applications in cloud computing, cost estimation, and optical networks.

Before concluding, it is worth highlighting another factor that distinguishes the inside and outside communities: *location*. To date, mainstream CBR venues have taken place in Europe and the US only, which may limit the interest of researchers from wider afield. This deserves further analysis than is possible here, but when we look at locations of the authors of these recent and influential papers we find, in a large majority of the cases, that the outside authors are based outside of Europe and the US; among the 15 outside papers, only 3 are from Europe or the US, with the rest from China, India, Korea, Pakistan, and Australia. This might be a sign that ICCBR needs to start looking further afield for future conference locations? It also points to a set of authors who may be well positioned to help organise and even host future ICCBR events.

6 Conclusions

The aim of this paper has been to examine the last three decades of CBR research. The data supports the existence of two significant but largely separate communities of CBR researchers: an *inside* community of AI/ML researchers, focusing on core techniques and applications, and an *outside* community of practitioners, focusing on a diverse tasks and applications from a variety of scientific and commercial domains.

The outside community is larger, but its members remain CBR authors for a shorter period of time. Many outside authors may be "passing through", leveraging CBR ideas in their research for a limited period of time only. The outside community's output lags behind that of the inside community in terms of citations, but this may be a consequence of the lack of connection between both communities and, the lack of citations from inside papers to outside papers, in particular. Despite this the outside community produces influential CBR papers, especially when we consider recent research.

One conclusion to draw from this is that it is worthwhile creating stronger links between both communities. Encouraging outside researchers to become involved in mainstream venues, may help to promote and sustain CBR within the outside community. Improving the flow of information between both communities will improve discoverability, especially for the work of the outside community. Furthermore, the outside community appears to be especially well positioned with respect to novel application domains for CBR, which may introduce new research challenges and themes to the inside community.

Precisely how we might bring about this increased engagement between the communities is a matter for the CBR community as a whole. There are some practical things that can be considered in the short-term, from inviting senior outside authors to present at ICCBR, to encouraging targeted sessions or workshops on emerging themes that are associated with the outside community. Longer-term actions might require other forms of outreach: involving outside researchers in ICCBR's programme and organising committees; encouraging host bids from locations that are well represented by the outside community (e.g. China, India, etc.). Whatever the approach, the good news is that, success will strengthen the field of CBR, helping to sustain the next 30 years of research.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun. **7**(1), 39–59 (1994)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
4. Ding, Y., Yan, E., Frazho, A., Caverlee, J.: Pagerank for ranking authors in co-citation networks. J. Am. Soc. Inf. Sci. Technol. **60**(11), 2229–2243 (2009)
5. Freyne, J., Coyle, L., Smyth, B., Cunningham, P.: Relative status of journal and conference publications in computer science. Commun. ACM **53**(11), 124–132 (2010)
6. Greene, D., Freyne, J., Smyth, B., Cunningham, P.: An analysis of research themes in the CBR conference literature. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 18–43. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85502-6_2
7. Greene, D., Freyne, J., Smyth, B., Cunningham, P.: An analysis of current trends in CBR research using multi-view clustering. AI Mag. **31**(2), 45–62 (2010)

8. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM (JACM) **46**(5), 604–632 (1999)
9. Kolodner, J.L.: An introduction to case-based reasoning. Artif. Intell. Rev. **6**(1), 3–34 (1992)
10. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
11. Slade, S.: Case-based reasoning: a research paradigm. AI Mag. **12**(1), 42 (1991)
12. Smyth, B., Keane, M.T.: Remembering to forget: a competence-preserving case deletion policy for case-based reasoning systems. In: Proceedings of the 14th International Joint Conference on Artificial intelligence (IJCAI), pp. 377–382. Morgan Kaufmann (1995)
13. Watson, I.: Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann Publishers Inc., Burlington (1998)
14. Wess, S., Althoff, K.-D., Richter, M.M. (eds.): EWCBR 1993. LNCS, vol. 837. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-58330-0