Assignment 2

Digital Libraries and Foundations of Information Retrieval

Winter semester 2022

1542011 Franka Brunen, 1365848 Andreas Schneider

Task 1: Author disambiguation in DBLP 15 Points

We disambiguate the profile of Arthur Schmidt (`https://dblp.uni-trier.de/pid/43/4520.html`):

(a) Arthur Schmidt, Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

Coauthor community: group 1

- conf/xsede/SatheesanABDGJK18
- journals/envsoft/ZimmerSOM15

Coauthor community: group 2

- journals/amco/ImbertJS10
- journals/jmc/EngelbertOS07
- journals/iacr/EngelbertOS06
- journals/moc/BuchmannS05
- conf/stoc/SchmidtV05
- journals/iacr/BuchmannGDELOSVW04

Coauthor community: group 3

- conf/fie/RoeslerLSSJMHMG15

(b) Arthur Schmidt, Darmstadt University of Technology, Germany

Coauthor community: group 4

- phd/de/Schmidt2007a

Task 2: Hirsch Index 3+4+4+4 Points

(a) If an author has an h-index of 11, it says that 11 of his publications have been cited at least 11 times. The h-index is always smaller than the number of publications.

paper	citation
1	213
2	201
3	180
4	155
5	122
6	99
7	88
8	67
9	56
10	25
11	18
12	11
13	9
14	6
15	3

(b) In this case, the h-index is not a good comparative measure because the scientists being compared are of different ages. This is a problem because the younger scientist logically has fewer publications to date than the older scientist. Accordingly, the younger one has less time to build up a high h-index.

Another criticism is that researchers in disciplines where there are fewer scientists may have a harder time getting to a high number of citations and therefore have a lower average h-index than researchers working in disciplines where there are very many other scientists (e.g., physics, biology). Therefore, in this case, the discipline must also be taken into account.

(c)
- Academic age: The academic age is the time that a scientist has been in the research field and performed active research. The academic age of a scientist may be computed as the span of years from their first published work up until the present. When the academic age is computed in formal settings, the academic age may be adjusted taking into account maternity and paternity leave, long-term illness, clinical training and/or national service.
- expert review: instead of citation, use experts to rate paper -¿ Importance difficult to measure right after publication -¿ can not be quantified.
- Altmetrics: paper reads/downloads, recommendations on social media,.. -¿ can be quantified, but easily manipulated.

(d) Field-weighted Citation Impact: FWCI equals to the total citations actually received divided by the total citations that would be expected based on the average of the considered field. FWCI of 1 means that the output performs just as expected for the global average. More than 1 means that the author outperforms the average, and less than 1 means that the author underperforms. One criticism of the h-index is that discipline is not included. The problem would be solved by using the industry-standard values, and comparisons could be made more effectively.

RA-Index: The RA-index accommodates improving the sensitivity of the h-index on the number of highly cited papers and has many cited paper and uncited paper under the h-core. This improvement can enhance the measurement sensitivity of the h-index. With the H-index, scientists with few but important publications are penalized. By giving more weight to highly cited papers, the RA-index compensates for this problem.

Task 3: Classification vs. full text 5+4+6 Points

- Database and storage security, information storage system, Information retrieval, information retrieval query processing, metrics, empirical studies, data management system
- With a keyword-based system, one advantage could be that the query is more efficient, as more precise searches are made for the keywords searched for. However, a disadvantage is that it is more time-consuming. In a keyword-based search, all content must be matched with the keyword, as opposed to classification.
- Searching based on classifications, for example, filters out all publications related to a specific topic. This could be a high number of publications if it is a broad topic. When searching for certain keywords in the titles of the publications, it is possible to search more specifically for certain topics. However, it may be that a publication does not appear in the search results, although it is about the searched topic, but just in the title the searched keywords do not appear. When searching for keywords in the full text of a publication, the entire text is searched, which means a high effort. Here it can be that many results appear.

Task 4: **Bonus task:** Computation of the Hirsch Index 15 Points

See file `dl-ir_2022ws_a02-t04_group2.py`.