Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 73 (2013) 544 – 550

Procedia
Social and Behavioral Sciences

The 2nd International Conference on Integrated Information

Institutional Name Confusion on Citation Indexes: The Example of the Names of Turkish Hospitals

Zehra Taşkın[a]* and Umut Al[a]

[a]*Hacettepe University, Department of Information Management, 06800 Beytepe, Ankara, TURKEY*

Abstract

Effectiveness of academia has usually been evaluated by using the number of publications. However, indexing with natural language on citation indexes makes evaluation process problematic for author affiliation. Affiliation information mistakes made by authors or editors can cause confusion during the indexing period. Consequently, visibility of the institutions in the citation indexes may reduce, different rankings for institutions may emerge and bibliometric studies may generate inaccurate results. It is important to understand these standardization problems in citation indexes to be able to fix them. The main aim of this study is to present name confusions of author affiliations within the citation indexes by using the example of the names of Turkish Training and Research Hospitals (TRH). In order to achieve this, 198,687 publications within the Web of Science, addressed in Turkey and produced between 1928 and 2009 were evaluated. The preliminary findings show that there are 65 unique TRHs which were mentioned with several different formats in the address section of the citation indexes. All these varying affiliation information were unified as a standard format for this study. It is observed that the names of these hospitals cause the main confusion. For example, there are four different "Numune TRH"s in different cities of Turkey. In such a case, if the author does not mention the city name in the address field, confusion becomes inevitable. In addition to that, affiliation information mistakes generally appear in the form of wrong spelling, abbreviation and translation mistakes. Findings of this study would bring up aforementioned confusion and present some recommendations to solve it.

Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information

Keywords: Name confusion; Web of Science; Citation indexes; Turkish Hospitals' names; standardization problem; unification of addresses.

* Corresponding author. Tel.: +90-312-2978200-130; fax: +90-312-2992014.
E-mail address: ztaskin@hacettepe.edu.tr

Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information
doi:10.1016/j.sbspro.2013.02.089

1. Introduction

Many institutions and universities are evaluated from the point of publication and citation count. However, quantitative evaluations are problematic because of some mistakes and data inconsistency on citation databases.

Citation databases are not only used to follow the literature, but also for citation analysis. Authors' and institutions' effectiveness have been measured by using citation analysis [1].

Parallel to the increase in the use of citation databases, their content grows rapidly. Therefore, some accuracy problems such as spelling, abbreviation and translation mistakes on the affiliation information [2] emerge. These mistakes can reduce institutional visibility and prevent collaboration opportunities for institutions. In addition to these problems, mistakes can also make performance evaluations unreliable.

For some countries (for example, Turkey) it is vital to access all the publications of institutions since these information is used for performance evaluation. However, data inconsistency should be taken into account during performance evaluation processes. The main aim of this study is to show affiliation information mistakes in the names of Turkish Training and Research Hospitals and to determine their effects. Some solutions will also be suggested at the end of the study.

2. Literature Review

Although the topic on data accuracy and consistency on citation indexes have drawn attention recently, there are only a few papers (e.g., [3], [4]) on this issue in the literature. These papers generally reveal the present condition and propose some solutions.

It is determined in a study that most of the mistakes are made in author and institution name fields [4]. The main inconsistencies for author names are namesake authors and different usages of author names. It is also found that using nicknames, changing surnames and translating Chinese, Korean and Russian author names into Latin alphabet create confusion about author names. According to this study, the confusion about address field is caused by the erroneous information provided by the authors. Two authors who work for the same institution can state two different addresses on their studies. Even the same author can state different addresses in different studies like "Hacettepe Univ, Dept of Information Management", "Hacettepe Univ, Dept of Information Studies", "Hacettepe Univ, Dept of Information and Records Management" etc.

De Bruin and Moed [5] emphasized that there can be difficulties in information retrieval because of unstandardized addresses. They suggested effective unification to solve these problems. Standardization problems may sometimes be based on institution names. Hood and Wilson [3] evaluated database usage on informetric studies and specified accuracy problems of databases. As a result, they found that the main accuracy problem was based on the changes of institution names. The names of universities, hospitals and corporations are changed frequently in Turkey. For example, Zübeyde Hanım TRH was founded as Ankara Maternity Hospital in 1960. Then the name of this hospital was changed into Social Security Ankara Maternity and Women's Health TRH, Etlik Maternity and Women's Health TRH, and finally Zübeyde Hanım TRH [6].

In addition to these works in the literature, there a few papers that have discussed the name confusion problems. Different techniques were used in these studies to identify and standardize data in the address fields.

The addresses of Turkish institutions have so far been evaluated merely from the point of standardization. A book was written by Turkish Scientific and Technological Research Center (TÜBİTAK) which included section on possible address variants of Turkish universities [7]. There is also a thesis about standardization of Turkish university names on citation indexes [8], which has explored the effects of standardization problems and presented a technique for unification by using finite state.

Only one study in the literature has dealt with the publications of training and research hospitals. A book published by TÜBİTAK presents the publication and citation counts of TRHs. However, this book seems to have

overlooked the possible address variants for these hospitals [9]. Nonetheless, this is the first study that reveals the institutional name confusion for TRHs in Turkey.

3. Methodology

First of all, we gathered 198,687 Turkey-addressed publications that were published between 1928 and 2009 and indexed in Web of Science (SCI, SSCI and A&HCI) by using the terms "Turkey", "Turkiye" or "Turkei" in the address field. There has been no publication type (article, proceeding or book review) restriction on our dataset. Address information for authors found in C1 and RP fields of Web of Science have been evaluated to identify institutional name confusions in citation indexes. A new column named "institution" has been created to write unified addresses for each institution by using Excel. For instance, if a publication has the address in C1 field like "DR ZEKAL TAHIR BURAK WOMEN HOSP, ANKARA, TURKEY; ZUBEYDE HANIM MATERN HOSP, ANKARA, TURKEY", the address of this publication has been written in the institution column as "ZEKAI TAHIR BURAK TRH; ZUBEYDE HANIM TRH". By this means, all the publications can be classified on the basis of their unified affiliation information.

After the unification process, publications written by TRHs were determined and saved in a different Excel file. Then, the most productive 20 hospitals and mistakes in their names were specified.

To present the effect of name confusion, bibliometric collaboration maps were created by using CiteSpace (http://cluster.cis.drexel.edu/~cchen/citespace/). Two collaboration maps were drawn to show the effects. First map included Web of Science affiliations (original addresses), and the second one used unified hospital names. Then, the differences between the two maps were evaluated. After presenting the adverse effects of confusion about hospital names, some suggestions were proposed at the end of the study.

4. Findings

A total of 198,687 papers were sorted by institution. After the sorting process, the most productive 10 hospitals and their loss on publications are listed in Table 1.

Table 1. Total publications of training and research hospitals and address mistake occurrences

Hospital	Pub. (N)	Mistakes (N)	Mistakes (%)
Ankara Numune TRH	2,325	437	18.7
Türkiye Yüksek İhtisas TRH	1,070	415	38.7
Ankara TRH	1,023	354	33.0
Şişli Etfal TRH	821	50	6.9
İzmir Atatürk TRH	648	493	76.0
Haydarpaşa Numune TRH	643	100	15.5
Dışkapı Yıldırım Beyazıt TRH	538	314	58.3
Dr. Siyami Ersek TRH	498	54	10.8
Dr. Abdurrahman Yurtaslan TRH	463	84	18.1
Dr. Sami Ulus TRH	444	10	2.2

The main problem was investigated especially in the case of hospitals with the same names. For example, there are two different "Atatürk Training and Research Hospital"s in two different cities in Turkey. One's name is "Ankara Atatürk TRH" and other's is "İzmir Atatürk TRH". If the author does not specify the affiliation address as Ankara or İzmir, the search results for these hospitals are really confusing. Therefore, the high error rate for İzmir Atatürk TRH is direct result of this confusion. The same problem is observed in the case of Numune Hospitals, too. Ankara Numune, Haydarpaşa Numune, Adana Numune and Trabzon Numune Hospitals have all been providing health services in Turkey for a long time. The name of Haydarpaşa Numune Hospital is grounds for further confusions as it resembles the name of Gülhane Military Medical Academy Haydarpaşa TRH

which, in turn, is easily confused with its parent institution, namely Gülhane Military Medical Academy. The names of these three hospitals make evaluation process highly challenging.

The other problem for hospitals is using place names instead of the proper names of the hospitals. For instance, authors show "Dışkapı TRH" as an affiliation. However, there are two training and research hospitals in Dışkapı district of Ankara, Turkey. When the authors do not specify their institutional affiliations by using the exact proper names, the ambiguity about addresses appears.

Fewer mistakes were determined for the hospitals which have uncommon and distinctive names like Şişli Etfal, Dr. Siyami Ersek and Dr. Sami Ulus Hospitals.

Error rates clearly show that evaluations for TRHs that depend on publication count must be implemented after deep data cleaning process. In the contrary case, the evaluation scores will be misleading. The negative effects of the name confusions will be discussed in the following part.

4.1. Effects of Name Confusion

Institutional name confusions have caused problems not only for performance evaluations, but also for governmental supports and academic studies. It is important to understand the negative effects of the problems to solve them efficiently.

TRHs are supported by TÜBİTAK in terms of gaining access to scientific information. Through a project named National Academic License for Electronic Resources, TÜBİTAK have purchased academic electronic resources for universities, TRHs and other academic institutions in Turkey [10]. The institutions to benefit from the project were chosen after an evaluation based on their scientific productivity by using their publication counts [7], [9]. The most productive institutions were more fortunate to meet the selection requirements than others. Therefore, accurate publication counts have become more important to be supported by governmental organization.

In addition to governmental support, affiliation information mistakes in citation databases make bibliometric studies meaningless and inconsistent. Along with the content development of citation databases, data accuracy and consistency have become more important for bibliometric studies.

Bibliometric mapping techniques have been used to show connections between authors, institutions or countries recently. Collaboration maps present relationships between institutions, so it is meaningful only if the institutional data is correct. According to this content, two maps were created for this study. First map has been created by deliberately using inaccurate data (see Fig. 1).

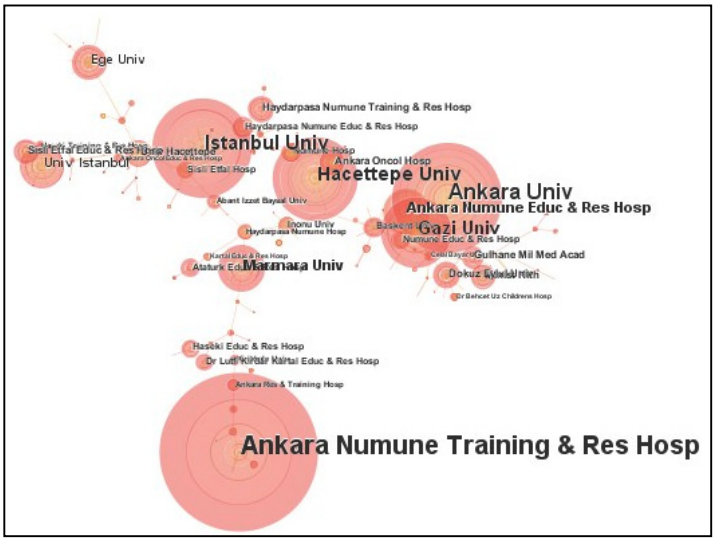

Fig. 1. Collaboration map with inaccurate affiliation information

Collaboration maps are important to show collaborative partners of an institution. However, as it is seen in Figure 1, the connections between institutions are really weak and cannot be determined effectively. It is also seen that the most productive hospital, Ankara Numune TRH, is shown on the map with its three different names. Due to this situation, commenting on the connections between organizations requires hard work.

Unified affiliation information have been used to create the second map (see Fig 2). It is obvious that the collaborations between institutions can be represented remarkably better.

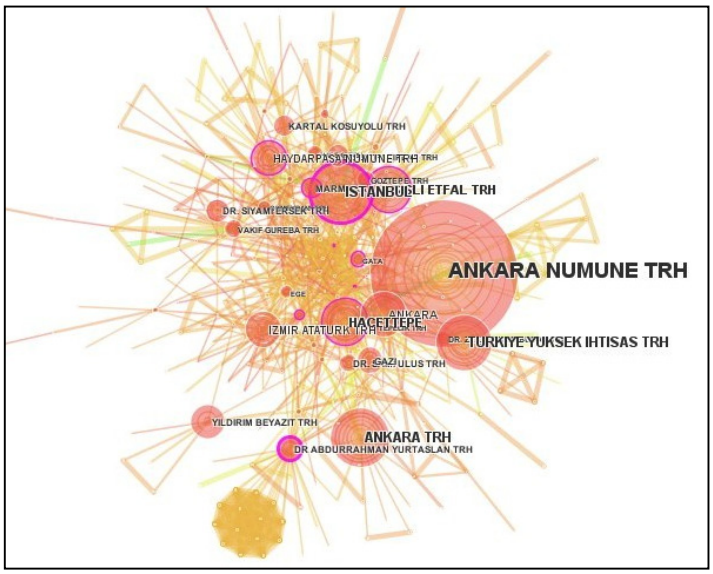

Fig. 2. Collaboration map with unified hospital names

The nodes that cannot be visualized in Figure 1, can easily be seen in the map of unified affiliations (Fig. 2). Figure 2 also shows the major knowledge-producer institutions and their connections with each other.

The difference between the two figures attests to the importance of well-structured unification process. However, working the unification process manually is time-consuming. If unification can be achieved by using automatic techniques, the results of bibliometric studies will be more effective.

4.2. Different Techniques for Unification

Although there are no studies related to the unification of hospital names, a few articles for university name unification, which mostly used clustering and finite state techniques, exist in the literature.

The unification process of clustering technique consists of cleaning, sorting, clustering, checking and updating stages. In this technique, incorrect addresses are determined by using a measure of similarity. In this way all strings are clustered and all possible errors are corrected [11].

The other technique, finite state, is used in the literature not only for standardization of author and institution names [12], [2], [13], [8]; but also for linguistic analysis, classifying genes, translating languages and etc. [14], [15], [16]. This technique depends on a finite set of states and their relations to each other [17]. A finite state automata accepts a string if it can follow a path from the beginning state to the final state [2]. Finite state transducers like Nooj or Xerox Finite State Tool implement this technique automatically to a dataset.

Universities and institutions can find out about the spelling, indexing and translating mistakes in their addresses by using the finite state technique. To access mistaken addresses on a dataset, a finite state graph should be drawn and implemented on all addresses. In a study which employed this technique, 49 different address variations were associated with Hacettepe University [8].

5. Conclusion

Organizations have been evaluated on the basis of their publication and citation counts recently. For this reason, performance evaluations of organizations depend on quantitative analysis. The main problem for evaluating by using citation databases is data inconsistency. There are many mistakes in institution names due to spelling, translation or indexing errors. Non-standardized addresses can reduce institutional visibility, different rankings and performances can emerge and bibliometric studies can produce unreliable results. Therefore, before the evaluation process, all existing institutional affiliation information must be unified.

Unifying addresses manually is a real challenge, so there are some techniques to make unification automatically. In the literature, clustering and finite state techniques have been used to determine and standardize mistaken addresses.

The main solution to the confusion about institutional names is to assign unique numbers to institutions. By using institutional IDs like DOI numbers, the confusion in institutional names can be minimized. Web of Science has a product named ResearcherID to give a unique number to authors. This project can be adapted for institutions. Scopus has already had a numbering system named affiliation identifier.

There are some responsibilities which authors, editors, librarians, indexers and decision-makers have to share in order to solve confusion problems. Authors must pay attention to write the correct addresses, editors must control these fields effectively. Librarians should lead authors to write down the correct addresses of their organizations. Decision-makers should consider data inconsistency in citation databases and prefer qualitative methods for evaluation instead of quantitative methods. Well-informed indexers should be employed and indexing should be made more carefully.

Acknowledgments

We thank Dr. Sinan Akıllı and Dr. İrem Soydal for their meticulous reading of a draft version of this paper and for their invaluable suggestions.

References

[1] Cole, J.R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In The Web of Knowledge: a festschrift in honor of Eugene Garfield, pp. 281-298. New Jersey: Information Today.
[2] Galvez, C. & Moya-Anegón, F. (2007a). Standardizing formats of corporate source data. Scientometrics, 70, 3-26.
[3] Hood, W.W. & Wilson, C.S. (2003). Informetric studies using databases: opportunities and challenges. Scientometrics, 58, 587-608.
[4] Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht: Springer.
[5] De Bruin, R.E. & Moed, H.F. (1990). The unification of addresses in scientific publications. Informetrics, 89-90, 65-78.
[6] Zübeyde Hanım TRH. (2010). Tarihçe (History). http://www.ezh.gov.tr/HastanemizHakkinda/Tarihce.aspx.
[7] ULAKBİM. (2007). Türkiye bilimsel yayın göstergeleri (Turkish National Science Indicators). Ed. İ.H. Demirel, C. Saraç & E.A. Gürses. Ankara: ULAKBİM.
[8] Taşkın, Z. (2012). Atıf dizinlerinde üniversite adreslerinin standardizasyon sorunu (Standardization problem of university addresses on citation indexes). Unpublished MA Thesis, Hacettepe University.
[9] ULAKBİM. (2008). Sağlık Bakanlığı Kurumlarının Türkiye'nin Bilimsel Yayın Sayısına Katkıları: 1981-2006 (Contribution of Ministry of Healths' organizations to the Turkish Scientific Publication Count). Ed. İ.H. Demirel, C. Saraç, E. Akıllı, Ö. Büyükçınar, S. Yetgin & E.A. Gürses. Ankara: ULAKBİM.
[10] TÜBİTAK. (2011). About TÜBİTAK ULAKBİM EKUAL. http://ekual.ulakbim.gov.tr/eng/about/.
[11] French, J.C., Powell, A.L. & Schulman, E. (2000). Using clustering strategies for creating authority files. Journal of the American Society for Information Science, 51, 774-786.
[12]Galvez, C. & Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). Scientometrics, 69, 323-345.
[13] Galvez, C. & Moya-Anegón, F. (2007b). Approximate personal name-matching through finite-state graphs. Journal of the American Society for In-formation Science and Technology, 58, 1-17.
[14] Oflazer, K. (1994). Two-level description of Turkish morphology. Literary and Linguistic Computing, 9, 137-148.
[15] Galvez, C. & Moya-Anegón, F. (2012). A dictionary-based approach to normalizing gene names in one domain of knowledge from the biomedical literature. Journal of Documentation, 68, 5-30.
[16] Özbek, G. & Jonathan, S. (2006). A suite of tools for augmenting English-to-Turkish statistical machine translation. Natural Language Processing Final Project http://infolab.stanford.edu/~jonsid/turkalator.pdf.
[17] Roche, E. & Schabes, Y. (1995). Deterministic part-of-speech tagging with finite-state transducers. Computational Linguistics, 21, 227-253.