eHealth2014 – Health Informatics Meets eHealth
A. Hörbst et al. (Eds.)

doi:10.3233/978-1-61499-397-1-47

Towards the Automated Generation of Expert Profiles for Rare Diseases through Bibliometric Analysis

Andreas PFLUGRAD[a,b,c,1], Karin JURKAT-ROTT[a,b], Frank LEHMANN-HORN[a,b] and Jochen BERNAUER[b,c]
[a] *Division of Neurophysiology, Ulm University*
[b] *Centre of Rare Diseases Ulm in the Centre of Excellence for Rare Diseases, Baden-Württemberg*
[c] *Computer Science Department, Ulm University of Applies Sciences*

Abstract. For patients suffering from rare diseases it is often hard to find an expert clinician. Existing registries rely on manual registration procedures and cannot easily be kept up to date. A prototype data collection system for discovering experts on rare diseases using MEDLINE has been successfully deployed. Initial manual analyses demonstrate proof of concept and deliver promising results. Examining the associations between authors, diseases and MeSH-Terms is expected to open up a variety of possibilities beyond expert discovery.

Keywords. Rare diseases, bibliometrics, information services.

1. Introduction

Nearly 7.000 different rare diseases have been described so far. With about 5% of the population suffering from a rare disease, patients often have to visit a multitude of physicians before finding the right expert clinician or centre of expertise to be correctly diagnosed and effectively treated. One problem for patients and physicians alike is a lack of information regarding experts on rare diseases. Therefore a need exists for dedicated expert registries.

Several possibilities on the internet assist the search for an expert. Orphanet, the largest European database for rare diseases and orphan drugs provides information regarding, amongst other things, centres of expertise, medical laboratories and patient organisations [1]. For discovering new experts, Orphanet as well as other registries are reliant on manual surveys, questionnaires and recommendations by known experts. Keeping information about experts, their expertise and their corresponding institution specific and up to date is hard to do and poses an immanent drawback of these procedures. The Centre of Rare Diseases Ulm is developing an automated system which employs bibliometric analyses to discover, retrieve and continuously update information on rare disease experts.

Similar approaches to expert discovery have previously been researched in several projects. Tang et al. [2] have implemented a researcher network knowledge base by

[1] Corresponding Author: andreas.pflugrad@uni-ulm.de

integrating publications from the Digital Bibliography & Library Project (DBLP) computer science bibliography as well as researcher web pages. Also, the agent based approach for finding experts within knowledge intensive organisations by Crowder et al. [3] and the semantic repository approach for locating academic experts proposed by Liu et al. [4] partly rely on publication analysis. One of the central premises of these approaches is that if a person has (co)authored a significant number of publications on a specific subject, this person can be seen as a potential expert in that subject [2, 5]. The project presented in this paper is based on the same premise.

Until now, the majority of projects and products for expert discovery regard companies and academic organisations, especially in the field of computer science [6]. These systems often include internal information such as e-mails. In the areas of medicine and biomedical research there is less experience with computerised approaches and a medical expert discovery system needs to be as transparent as possible. Therefore, the project at hand is based on publicly available, verifiable information and involves a revision by medical domain experts.

2. Methods

2.1. Project Overview

The project aims at developing a computerised system for discovering experts on rare diseases. For that purpose, expert profiles are automatically generated and maintained by analysing MEDLINE [7]. The integration possibilities of other information sources including guideline repositories [8], research networks [9] and research projects, e.g. those funded by the German Research Foundation, will be investigated at a later stage. As a first step, a system has been developed, which collects meta-data of articles on rare diseases by utilising the PubMed search application programming interface (API). Authors of retrieved articles are linked to the particular disease entity and all associated Medical Subject Headings (MeSH).

The early stage of profile generation and analysis focuses on the publication count of each recorded author whilst considering the respective position as first, middle or last (senior) author. The publication count can be seen as a measure of an author's involvement in a specific disease entity. The MeSH-Terms attributed to a specific author are the basis for (1) determining the author’s field of expertise and (2) categorising the author’s professional focus such as basic research, diagnostics, or therapy. In addition, it will be examined, how further analysis of the data can provide useful information about experts and diseases. Eventually, the profiles are to be refined, extended and made available to professionals and patients. Profilers of manually maintained expert registries could adapt the data to discover new experts, gain additional analysis possibilities or register institution changes which might have gone unnoticed otherwise thus enhancing the quality of their data.

2.2. Preliminary Work

For the initial data collection, the freely accessible MEDLINE database has been chosen. It comprises more than 21 million records from biomedicine and health subjects and is likely to cover most relevant publications on rare diseases. A thesaurus has been created from the Orphadata rare diseases directory in combination with other terminologies containing synonyms and classification relations such as MeSH or the Online Mendelian Inheritance in Man (OMIM) database [10]. The thesaurus serves as a reference database for querying PubMed and contains multiple terms for each disease entity, using the Orpha-Number (OrphaNo) as a common unique identifier. After eliminating potential redundancies, first tests showed several composite terms used by Orphanet. When used in a query, erroneous or no results were received from PubMed. After splitting the terms into their basic forms, more feasible results could be obtained. At the time of this paper, the reference database contains 6.771 OrphaNos with a total of 27.198 query terms. The number of query terms per OrphaNo ranges from 1 to 18 with an average of 4.

2.3. Search Strategies

A suitable search strategy had to be found in order to retrieve all relevant articles from PubMed and minimise the number of irrelevant ones. It was decided to use the "Title" and "MeSH Major Topic" fields in conjunction with a logical OR for each term of a disease entity to find the relevant articles. Using a logical AND or the Major Topic field alone turned out to be too restrictive while using only the Title field yields more results but may still leave out relevant articles.

Early tests with selected disorders showed that several relevant articles were not captured due to a lack of query terms in the initial thesaurus. Therefore, a graphical user interface has been designed which allows domain experts to add terms for a disease entity. Existing terms which have been found to not yield any results can be modified or deleted. This way it is possible to obtain a comprehensive and adaptable data basis which contains only relevant query terms.

2.4. Data Collection

A program was developed to automatically search and retrieve MEDLINE publication data via the PubMed API for each disease term listed in the thesaurus. This is done by using the National Center for Biotechnology Information's E-utility web service. All retrieved data is fed into a staging database which serves as the starting point of profile generation.

Scaling methods were developed to increase the performance of the program in retrieving a large amount of MEDLINE data. The search results of a PubMed query are split up into packages of 500 articles each. Using bigger packages resulted in erroneous responses while smaller packages led to a performance decrease. Fetching an article package may still result in a program exception when corrupt data is returned for a single or several article positions. This is handled by splitting up the afflicted package into smaller sub-packages which are then fetched again. This procedure is reiterated until all retrievable articles of the package have been captured.

If a previously captured article is found in a subsequent search run, the program checks its publication status. If it has changed, indicating revised or added data, e.g. through MEDLINE indexation or the addition of MeSH-Headings, the article data is captured again and changed respectively.

2.5. Profile Generation

Sample profiles have been generated and allow for a first analysis of article meta-data regarding the number of publications of an author as well as the interpretation of the associated MeSH-Terms. This preliminary analysis serves as a proof of concept of the overall approach before more sophisticated methods for profile generation, enrichment and analysis are employed. All author entries with the same last name and initials were grouped into a single author entity and the relevant data was aggregated for each entity. The first name was left out due to inconsistent occurrence. This naive approach to profile generation is prone to distortions caused by name ambiguity i.e. different authors sharing the same name as well as name variations of the same author.

2.6. Data Analysis

The first metric to be looked upon with the sample profiles is the article count of an author including the frequency of being in first, middle or senior position. The second metric regards the MeSH Descriptors and Qualifiers which are associated with a specific profile and can be used to determine an author's field of expertise. Two kinds of MeSH-Descriptors can be discerned: (1) subject topics such as diseases, chemicals or procedures and (2) subsidiary topics denoting age groups, study types or living subjects. Both types allow for a description of an author's field of study. Additionally, the MeSH-Qualifiers can be used to specify the principal activity such as basic or clinical research.

Key figures of the staging database have been collected and are represented in section 3.1. The analyses include article counts for different OrphaNos and authors as well as the number of MeSH-Terms for different diseases, articles and authors. The results are used to assess the quality of the data as well as to detect potential pitfalls and improvement possibilities for the data collection process. Sample profiles of a particular author and a disease entity within that author's expertise in section 3.2 illustrate the data relevant for determining and validating an author's field of expertise.

3. Results

3.1. Staging Database

At the time of this paper (January 2014) the initial data collection has been conducted for roughly one third of the disease entities in the reference database. Over the course of just over 10.000 PubMed queries, more than one million articles and over 5 million authorships, which can be assigned to over 1.3 million authors, have been retrieved from MEDLINE. The exact numbers are shown in Table 1.

Fig. 1 shows how many articles and authors have been retrieved for how many of the covered disease entities. The left-hand part shows the amount of OrphaNos for which a certain number of articles has been retrieved. It can be seen that for 1.017

Table 1. Key figures of the staging database. *The number of distinguishable author entries underlies the restrictions of the preliminary aggregation approach described in section 2.5.

Processed disease entities	Conducted PubMed queries	Retrieved Articles	Authorships	Authors*
2.606	10.368	1.259.751	5.438.607	1.306.714

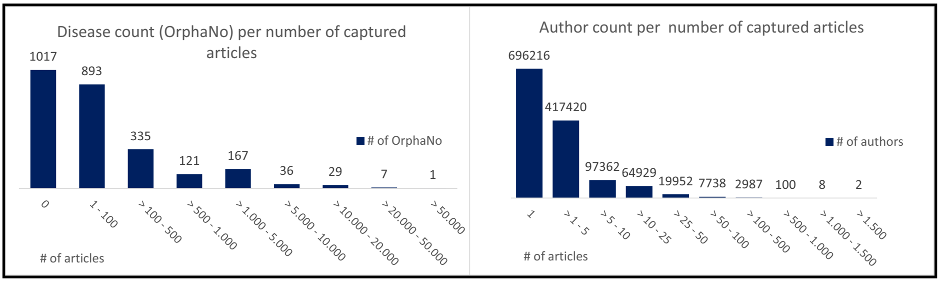


Figure 1. Disease and author count per number of captured articles

disease entities no articles could be retrieved from PubMed. While this may, for some OrphaNos, be attributed to an actual lack of publications in MEDLINE, inadequate query terms in the reference database are likely to be the main issue. It can be noted that for 893 OrphaNos at least one and at most 100 articles could be retrieved. Queries for OrphaNo 543 (Burkitt lymphoma) resulted in more than 50.000 retrievable articles.

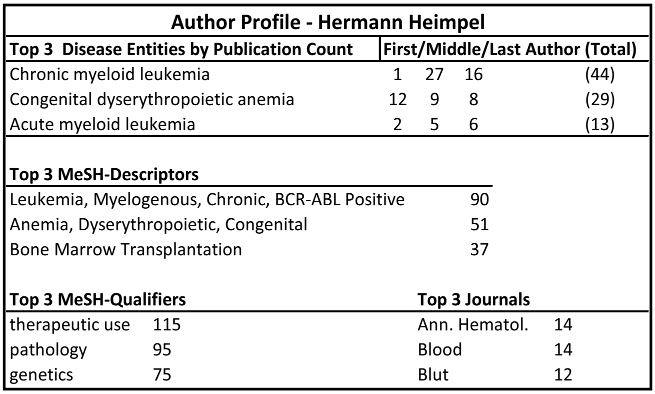
The right-hand side of Fig. 1 displays, how many authors can be allocated to a group of retrieved articles. It can be seen that with 696.216, more than 50% of the authors in the staging database are registered with only a single article. This particularly high number can partly be attributed to spelling differences in the name of the same author across different articles. Another 417.420 (30%) of the retrieved authors are registered with up to 5 articles. Exceptionally high publication counts occur for names such as J. Zhang and W. Wang, which are used as a benchmark in disambiguation research [11]. It is, however, not unfeasible for single researchers to have more than 500 publications. An example from the current data set is RJ Wanders, whose high publication count can be confirmed by his ResearchGate profile [9].

Table 2 shows associations between different key figures. The first column shows how many different disease entities are associated with a single author. The mean value of two articles per author corresponds to the distribution seen in Fig. 1. The maximum number of 335 diseases for a single author may again be an ambiguity issue. The low mean and median values show that the scope of expertise is rather narrow, indicating that the experts are highly specialised for very few rare diseases.

Columns two to four show the associations of the stored MeSH-Descriptors per disease entity, author and article. The minimum value of zero corresponds to retrieved articles which have not or not yet been indexed for MEDLINE. These numbers are to be complemented over the course of future search runs. Disease entities are associated with a mean value of 732 descriptors, peaking at 11.205 for Burkitt lymphoma. The

Table 2. Aggregated key figures of associations between the collected data in the staging database. *The number of distinguishable author entries underlies the restrictions of the preliminary aggregation approach described in section 2.5.

	OrphaNos per author*	MeSH-Descriptors per OrphaNo	MeSH-Descriptors per author*	MeSH-Descriptors per article
Minimum	1	0	0	0
Maximum	335	11.205	4.142	103
Average	2	732	30	10
Median	1	199	16	10



Author Profile - Hermann Heimpel

Top 3 Disease Entities by Publication Count	First/Middle/Last Author (Total)			
Chronic myeloid leukemia	1	27	16	(44)
Congenital dyserythropoietic anemia	12	9	8	(29)
Acute myeloid leukemia	2	5	6	(13)

Top 3 MeSH-Descriptors	
Leukemia, Myelogenous, Chronic, BCR-ABL Positive	90
Anemia, Dyserythropoietic, Congenital	51
Bone Marrow Transplantation	37

Top 3 MeSH-Qualifiers		Top 3 Journals	
therapeutic use	115	Ann. Hematol.	14
pathology	95	Blood	14
genetics	75	Blut	12

Figure 2. Sample profile of a particular author.

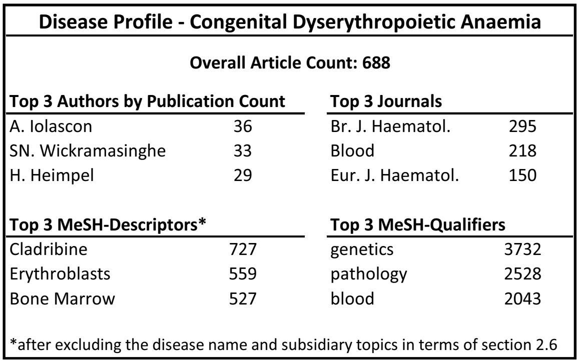
median value of 199 reflects the high number of articles for certain disease entities, resulting in an increased average value. An author is on average attributed with 30 descriptors although this number is likely to be distorted by the high maximum value which is again affected by the lack of name disambiguation. With a median value of 16, the analysis of MeSH-descriptors for single authors is considered eligible. Retrieved articles are indexed with mean and median values of ten descriptors with a maximum of 103.

3.2. Sample Profiles

Fig. 2 shows a sample profile of a particular author (H. Heimpel) including the top three entries for the author's publication count, MeSH-Terms and journals. The author's most numerous articles are on the subjects of acute or chronic myeloid leukaemia as well as congenital dyserythropoietic anaemia (CDA).

The predominant MeSH-Descriptors in the expert profile are "Leukemia, Myelogenous, Chronic BCR-ABL Positive" and "Anemia, Dyserytropoietic, Congenital". Additionally, "Bone Marrow Transplantation" is denoted as the author's most frequently described procedure. The Qualifiers show an emphasis on therapy and pathology followed by genetics. All of the top three journals are on the subject of haematology. Summarising the profile information, the author could be classified as a clinical haematologist and oncologist who specialises in rare leukaemia and anaemia with a particular involvement in CDA, which is indicated by the high number of publications as first author.

Fig. 3 shows a sample profile of a disease entity. Similarly to the author profile shown in Fig. 2, the disease profile includes the top three MeSH-Descriptors, Qualifiers and journals. Additionally, the authors with the highest publication count for this disease are listed. Several connections between the presented sample profiles can be seen. The author depicted in Fig. 2 is ranked among the top three publishing authors for the presented disease. Further commonalities include the journal Blood, a MeSH-Descriptor regarding bone marrow as well as the Qualifiers genetics and pathology.



Disease Profile - Congenital Dyserythropoietic Anaemia

Overall Article Count: 688

Top 3 Authors by Publication Count		**Top 3 Journals**	
A. Iolascon	36	Br. J. Haematol.	295
SN. Wickramasinghe	33	Blood	218
H. Heimpel	29	Eur. J. Haematol.	150

Top 3 MeSH-Descriptors*		**Top 3 MeSH-Qualifiers**	
Cladribine	727	genetics	3732
Erythroblasts	559	pathology	2528
Bone Marrow	527	blood	2043

*after excluding the disease name and subsidiary topics in terms of section 2.6

Figure 3. Sample profile of the disease entity CDA.

4. Discussion

After querying PubMed for one third of the disease entities in the reference database, a considerable amount of data has been successfully retrieved. For an unexpectedly high number of disease entities, no articles could be obtained, indicating a need for improving the reference database and the search strategy, e.g. by also screening the abstract field and using query terms that include superordinate and generic terms for disease groups. Additionally, further query terms for individual disease entities may have to be recommended by domain experts. In order to further enhance the completeness of the profile database, other data sources and media, such as biomedical textbooks and guidelines, will be included. Furthermore, comparisons with the expertise documented in the ResearchGate database [9] as well as in therapeutic guidelines databases such as [8] will be investigated as a means to differentiate ambiguous experts and validate individual profiles and publication counts. Whether these can be incorporated in an automated manner is subject to future research. Distortions arising from the simplistic grouping approach described in section 2.5 showed the necessity of sophisticated disambiguation and harmonisation methods to be employed for the profile generation process.

The preliminary database has been useful in the Rare Diseases Centre Ulm for proper patient referral to an appropriate expert. However, it must be noted that bibliometric analysis can only provide insights on rare disease experts who actively publish. Experts with no publishing activity are unlikely to be discovered. Additionally, not every author may be an actual expert on the research topic, e.g. in the case of honorary authorships. The approach may then still point to the right department, if not necessarily to the right person. This may be remedied by tracking the authorship position and weighing accordingly.

Overall, the first analysis of the project has been positive. The sample profiles presented in section 3.2 have shown that the data can be used to create a sufficiently accurate determination of an author's field of expertise. It provides a foundation for more detailed analysis and validation to be carried out for selected rare diseases within the expertise of the Centre of Rare Diseases Ulm. From these first steps within the narrow scope of MEDLINE analysis, the system can be employed to support manual profiling by identifying new experts and providing additional insights to existing ones. It may then gradually be extended, e.g. by using web crawling mechanisms for obtaining further data such as contact information of experts. While patients increasingly utilise social media and special interest groups to find help and exchange views, it might be helpful to provide an additional source of impartial information. Ultimately, the system will enable the advice seeking patient to contact the right expert or institution. At this stage of development it is too early to evaluate how the presented approach will perform in comparison to these alternatives. It is, however, the first automated expert retrieval system on the subject of rare diseases providing transparent and verifiable information. With further development and refinement, the system may also be adapted to suit a broader medical field.

Acknowledgements

The authors wish to thank the Ministry of Science, Research and the Arts of Baden-Württemberg, the non-profit Eva-Luise Köhler and Hertie Foundations, the German Federal Ministry of Research (IonNeurONet) and the Friedrich-Wingert Foundation for supporting this project.

References

[1] Orphanet: an online rare disease and orphan drug data base. Copyright, INSERM 1997. Available on http://www.orpha.net, last access 25.1.2014

[2] Tang, Jie, et al. "ArnetMiner: An Expertise Oriented Search System for Web Community." *Semantic Web Challenge*. 2007.

[3] Crowder, Richard, Gareth Hughes, and Wendy Hall. "An agent based approach to finding expertise." *Practical Aspects of Knowledge Management*. Springer Berlin Heidelberg, 2002. 179-188.

[4] Liu, Ping, Yan Ye, and Kan Liu. "Building a Semantic Repository of Academic Experts." *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on*. IEEE, 2008.

[5] Stankovic, Milan, et al. "Looking for experts? What can linked data do for you?" *LDOW*. 2010.

[6] McDonald, David W., and Mark S. Ackerman. "Just talk to me: a field study of expertise location." *Proceedings of the 1998 ACM conference on Computer supported cooperative work*. ACM, 1998.

[7] PubMed. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed, last access: 17.3.2014

[8] AWMF-online Leitlinien, http://www.awmf.org/leitlinien.html, last access: 27.1.2014

[9] ResearchGate, http://www.researchgate.net, last access: 17.3.2014

[10] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), http://omim.org/, last access: 26.3.2014

[11] Tang, Jie, et al. "A unified probabilistic framework for name disambiguation in digital library." *Knowledge and Data Engineering, IEEE Transactions on* 24.6 (2012): 975-987.