LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 31, (2021) 1–22 | e-ISSN: 2213-056X

Words Algorithm Collection – Finding Closely Related Open Access Books using Text Mining Techniques

Ronald Snijder

OAPEN Foundation, The Hague, Netherlands
r.snijder@oapen.org, orcid.org/0000-0001-9260-4941

Abstract

Open access platforms and retail websites are both trying to present the most relevant offerings to their patrons. Retail websites deploy recommender systems that collect data about their customers. These systems are successful but intrude on privacy. As an alternative, this paper presents an algorithm that uses text mining techniques to find the most important themes of an open access book or chapter. By locating other publications that share one or more of these themes, it is possible to recommend closely related books or chapters. The algorithm splits the full text in trigrams. It removes all trigrams containing words that are commonly used in everyday language and in (open access) book publishing. The most occurring remaining trigrams are distinctive to the publication and indicate the themes of the book. The next step is finding publications that share one or more of the trigrams. The strength of the connection can be measured by counting – and ranking – the number of shared trigrams. The algorithm was used to find connections between 10,997 titles: 67% in English, 29% in German and 6% in Dutch or a combination of languages. The algorithm is able to find connected books across languages. It is possible to use the algorithm for several use cases, not just recommender systems. Creating benchmarks for publishers or creating a collection of connected titles for libraries are other possibilities. Apart from the OAPEN Library, the algorithm can be applied to other collections of open access books or even open access journal

openjournals.nl | http://liberquarterly.eu/ | DOI: 10.53377/lq.10938

articles. Combining the results across multiple collections will enhance its effectiveness.

Keywords: Open access; recommendations; books; algorithms; text mining

1. Introduction

Open access platforms and retail websites have one thing in common: they are trying to present the most relevant offerings possible to their patrons. Retail websites – such as Amazon.com – deploy recommender systems based on data collected about their customers. These systems improve with the amount of data available: the more is known about the customers, the better it can predict what other merchandise will appeal.

For open access platforms, this is not a viable solution. First, these platforms are designed to lower as many barriers as possible to make sure that the largest group of people have access to the publications. Forcing people to identify themselves and tracking their actions on the website is a serious barrier. Second, and more importantly, protecting privacy is an important principle in the library community which is at the very least overlapping with the open access community.

Recommender systems are successful but using open access platforms to track people is not acceptable. Therefore, a different solution is needed. Compared to retail websites, open access platforms have a unique advantage: they are able to use the complete contents of the publications they host. So, the question arises if it is possible to create a recommender system based on the contents of freely available documents, instead of personal data. This paper presents an algorithm that uses text mining techniques to find the most important themes of an open access book or chapter. By locating other publications that share one or more of these themes, it is possible to recommend closely related books or chapters.

The algorithm splits the full text of the book or chapter in sets of three consecutive words: trigrams. Then it removes all trigrams containing words that are commonly used in everyday language and the trigrams containing terms that are commonly used in (open access) book publishing. When a trigram contains a word – or multiple words – that is commonly used, the whole trigram is discarded.

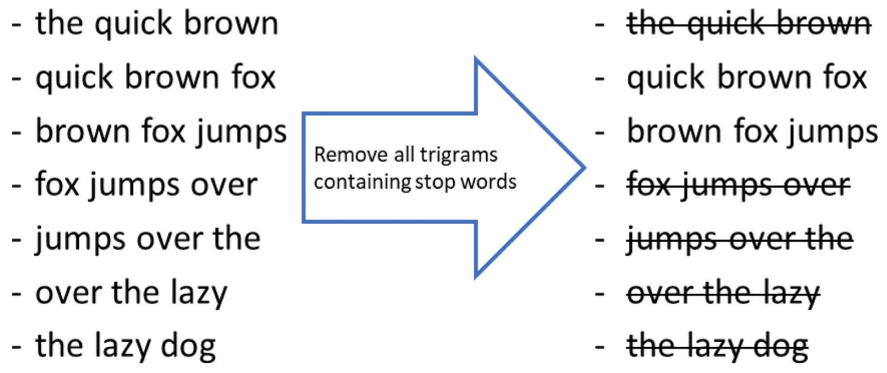
Figure 1 illustrates this using a simple sentence: "The quick brown fox jumps over the lazy dog". Converting the sentence to trigrams results in seven sets of three words, the trigrams. Removing all trigrams that contain commonly used words brings the remaining number back to two. Deploying this procedure to the complete text of a book still creates a large set of trigrams, hence the need for additional filtering using terms that are common for open access academic books.

The remaining trigrams are distinctive to the book or chapter and selecting the most occurring of those trigrams indicates the concepts the author of this title is discussing. The next step is finding publications that share one or more of the trigrams; the more trigrams they share, the closer the connection between them. The strength of the connection can be measured by simply counting the number of shared trigrams.

In contrast to black box technologies such as machine learning[1], the algorithm is completely transparent. Every term used is open to scrutiny and can be updated. Furthermore, the algorithm is tool agnostic: it is not tied to a specific coding environment.

The solution described in this paper is based on standard open-source software. It is built using a combination of DSpace 6 and the R programming language. The open access platform – based on DSpace 6 – is the OAPEN Library; the data set used consists of nearly 11,000 open access books and

Fig. 1: Trigrams example.



chapters. The OAPEN Library enables data extraction through an API (application programming interface). A text mining algorithm written in the R programming language uses the full text of the publications, filters out the trigrams and creates an overview of closely related books and chapters.

Different users may have different needs: a reader might be interested in finding a few select titles, while a library might want to download a larger collection of books around a certain topic. These use cases are discussed in section 4.4.

2. Background

As mentioned in the previous section, the set of publications is provided by the OAPEN Library. The OAPEN Library is a platform – launched in 2010 – hosting open access books and chapters. It is managed by the OAPEN Foundation[2]. In June 2021, the collection consists of over 17,000 titles. This background section discusses privacy in libraries, recommender systems, ngrams and previous experiments run on the OAPEN collection.

2.1. Libraries and Privacy

Libraries – whether physical or online – have been protecting the privacy of their patrons for quite some time; for instance by the American Library Association (ALA) Code of Ethics in 1938 (Witt, 2017). This position is shared among the International Federation of Library Associations and Institutions (2016), the American Library Association (2014) and several other national library associations. Privacy in libraries is associated with protection from unwanted government attention (Jaeger et al., 2004), but also from commercial organisations (Corrado, 2007; Maceli, 2018).

2.2. Recommender Systems

Recommender systems are used to provide suggestions about items that are valuable to a person. While there are several techniques for building recommender systems, most are based on the same principle: create a profile of the

user and her peers, extend this as much as possible and update it over time. This enables the system to know the preferences of the user and thus predict other items (Pazzani & Billsus, 2007; Ricci et al., 2011; Schafer et al., 1999). Linden et al. (2003) and Smith and Linden (2017) discuss their experiences at Amazon, spanning two decades.

For those who do not feel comfortable with the lack of privacy in connection to these type of systems, Jeckmans et al. (2013) have listed countermeasures. These include raising awareness about privacy issues and invoking specific laws dealing with personal information. As it might take quite some time before this will take effect, the authors also describe technical measures such as anonymisation, randomisation and the use of cryptography.

Instead of recommending titles based on personal data, here the contents of the titles will be used. The texts of the books and chapters are analysed using ngrams.

2.3. Ngrams

Ngrams are based on the relationships between words, either by examining which words tend to follow others immediately, or by looking at words that co-occur within the same documents. Two consecutive words are called "bigrams", three consecutive words are called "trigrams". Naturally, the number of trigrams in a text is lower compared to bigrams, while the trigrams are more specific. As we are examining a large text corpus – the text of almost 11,000 books and chapters – the total number of possible trigrams is still large.

Ngrams are used in different types of research. One application is document clustering: creating related groups of documents. Each document is represented by a numerical value. The k-means algorithm is typically used to calculate the distance between the documents and a 'cluster means'; the goal is to all documents in clusters with the smallest numerical distance (Miao et al., 2005). Furthermore, the authors looked at the performance of several types of ngrams – ranging from bigrams to 5-grams – used in document clustering. They conclude that trigrams are roughly as accurate as 4-grams and 5-grams but are more economical in their resource usage.

Apart from clustering documents, ngrams are also deployed for author attribution. This technique aims to find the characteristics of a writer's style and use that to define whether a certain text is written by that author. Here, the ngrams are not based on clusters of words, but clusters of characters (Kešelj et al., 2003). Eve (2019) is critical of the application of this technique to identify authors and uses it to distinguish literary genres instead.

The best-known use of ngrams is probably the Google Books Ngram Viewer. This vast corpus of books is used for cultural research. The most cited example is written by Michel et al. (2011). In this paper, the authors examine the change of language over time, but also cultural changes: the rise and demise of the celebrity of certain persons and suppression of ideas over time. This is far from the only paper based on the Google Books Ngram Viewer: a recent search on this subject in the Google Scholar search engine resulted in over 3,700 titles[3].

The experiment of this paper does not quite fit within these three research types. It is clearly not meant to discover long term trends, in the manner of the Google Books research. Finding authors is also not necessary: this information is provided in the metadata of the OAPEN Library. The k-means algorithm is a more general-purpose application, aimed to be useful in various situations.

The most closely connected experiments are those aiming to extract keywords from an article or book (Rohini & Ambati, 2007; Souza & Raghavan, 2014). The authors describe the use of statistical methods to find distinctive words. However, the text corpora used are small and no attempt is made to connect multiple titles.

The algorithm used in this experiment is optimised for a very specific purpose: instead of creating amorph groups, it aims to find exact relations for each individual title. These relations – based on the number of shared trigrams – are ranked. The ranking and the number of shared trigrams can be used to create services for digital libraries with an open access collection. This algorithm is not general-purpose but optimised for one specific environment.

2.4. Other Experiments

Several other experiments have been conducted on the OAPEN Library collection: creating groups of books based on usage data (Snijder, 2019) or

categorising titles based on Wikipedia pages (Snijder, 2021). In the first experiment, the download patterns are analysed to find which books are regularly selected together. So, instead of looking at individual preferences, social network analysis was deployed to find the preferences of groups of people. The more recent investigation aimed to categorize books by automatically finding the Wikipedia pages that describe their contents.

Grouping books based on usage data has drawbacks: apart from the reliance on external usage data, the results need to be interpreted. The interpretation depends on analysing aspects of the books and the users. This cannot be automated, making it hard to upscale, and the analysis might be open to bias. Furthermore, using data captured on different time periods lead to different results.

Another way to discover similar titles is by adding standardised metadata. Most libraries use a classification for this purpose, which is standardised but rigid. Another option is using uncontrolled keywords that are flexible but lack standardisation. Wikipedia was used as 'middle ground': a standardised but very broad set of keywords. Adding Wikipedia pages to book records in the OAPEN Library is also reliant on external data, which must be provided by separate service. Furthermore, manual 'culling' of the results was necessary.

Both methods cannot be implemented completely automatically, rely on external services and need extra effort to scale up. This makes them less desirable for production. The solution described in this paper does not rely on external services but uses the strength of open access publishing: direct access to the contents of the documents.

3. Finding Related Titles by Algorithm

This section describes the algorithm used and the data set. The text mining techniques deployed are built using the work by Silge and Robinson (2016, 2017). The authors created a set of tools ("package") in the programming language R (R Core Team, 2020) aimed to simplify text mining. The R package creates the trigrams, which are manipulated to find the related documents.

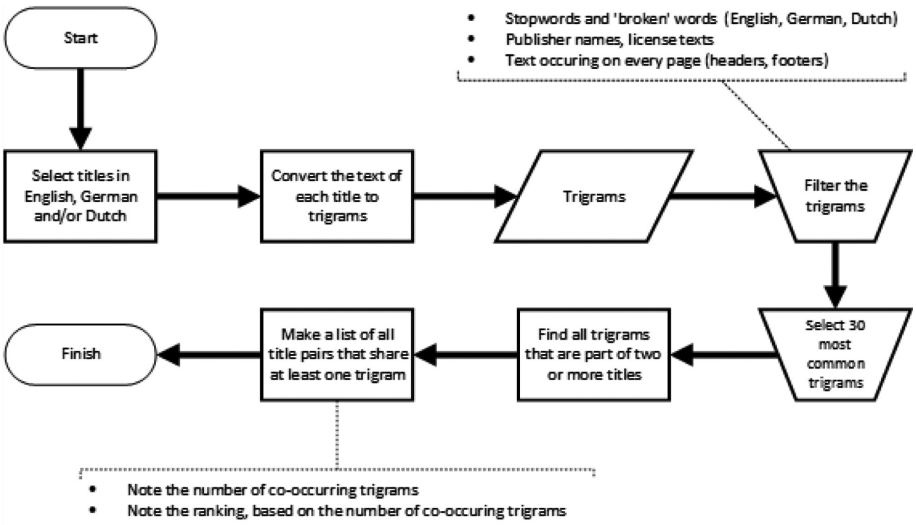
3.1. The Algorithm

Our goal is to find relevant open access titles, when a book or chapter has been selected. Relevant titles discuss the same concept or concepts that are closely connected. The algorithm is based on two assumptions: 1. The terms describing the themes of the title are frequently occurring in the text; 2. Books and chapters on the same subject use similar terms. In other words: if titles share relevant terms, they are connected. The number of shared relevant terms is an indication of the strength of the connection. Figure 2 displays the complete algorithm.

The next question is what terms to use. In this experiment, the terms are sets of three words – trigrams. In a text, the number of trigrams is relatively small – compared to bigrams – while they are more specific. This leads to a more 'workable' set of possible items. However, not all trigrams are relevant for our purpose, and therefore it is important to filter out the ones that are not needed.

The first set of trigrams to discard contains words that are too common: stop words. Examples are "a"; "able"; "about"; "above", and almost 1,200 more

Fig. 2: The algorithm as flow chart.



words for the English language. Comparable sets of stop words for German and Dutch were also deployed.

The next set to filter out is trigrams that contain parts of words. When the contents of the books are converted to text, hyphens are converted to spaces, leading to trigrams such as "diff ere nt", "inso fe rn" or "werkge legenhe id". These are not three words, but just one.

Furthermore, trigrams that are specific to open access publishing or academic writing are discarded. These are descriptions of Creative Commons licenses, or terms that are quite common in academic books, but are meaningless in themselves, such as "pdf letzter zugriff", "pdf zuletzt geprüft", phd diss university" or "phd thesis university".

Also, the part of references that only contain the publisher's name are filtered out. For instance, the trigram "manchester university press" does not convey which title is cited. As Manchester University Press has published hundreds of titles on many different subjects, linking books using this term does not describe any subject related connection. Of course, this also applies many other academic publishers.

It is important to note that the terms to be excluded are a clearly visible part of the algorithm. This ensures maximum transparency: each person working with the algorithm has direct access to the 'filtering terms' and might choose to update them.

3.2. The Data Set

At the start, 12,224 titles in the OAPEN Library were selected. The selection was based on one criterium: language. The books and chapters were published in English, German, Dutch or a combination of these languages. Choosing these three languages was pragmatic: over 90 percent of the OAPEN Library collection is published in either English, German or Dutch, ensuring a sizable set of titles to analyse. Having a data set spanning multiple languages also enables possible connections between books in several languages. In one of the examples in section 4.2, we will find two closely connected books: an English language translation of a German book.

The first phase of data gathering was an attempt to download the full text of the titles. From each text, the most relevant trigrams were selected and lastly, for each title was determined if it could be matched to one of more other books or chapters. During this process, some texts could not be extracted, or no matching title could be found. This led to a dropout rate of around 10 percent, resulting in the 10,997 titles of the data set.

The data set is dominated by books (see Table 1); only 4% are chapters. Within the data set, English and German stand out, with a small percentage – 6% – of titles in Dutch or in multiple languages.

Each book or chapter in the data set is connected to one or more titles. The majority of the titles – over 7,000 – are closely related to 50 titles or less. Another 1,986 are connected to 100 titles or less. When the largest group is subdivided, it becomes clear that 4,498 books or chapters are closely connected to 20 titles of less. In other words, 40% of the titles.

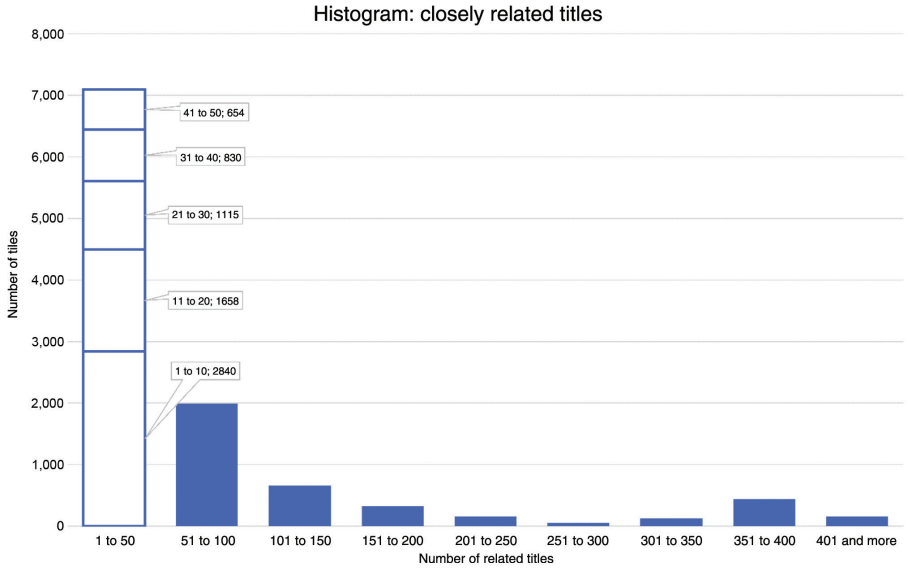
Each title shares one or more trigrams with another publication. As is clearly visible in the histogram (Figure 3), most books and chapters are connected to 21 titles or more. Most of these connections vary in the number of shared trigrams. The number of shared trigrams is an indication of the strength of the connection: a higher number indicates a stronger connection.

These connections could be ranked. For instance, if a book is connected to 25 books – two books with three shared trigrams, five books with two shared trigrams and the rest with one shared trigram – these could be ranked first,

Table 1: Publication types and languages.

Publication type	Language	Amount	Percentage
Book	English	6,736	61%
	German	3,155	29%
	Dutch	521	5%
	Multiple languages	97	1%
Total book		**10,509**	**96%**
Chapter	English	467	4%
	Dutch	9	0%
	German	9	0%
	Multiple languages	3	0%
Total chapter		**488**	**4%**
Total		**10,997**	**100%**

Fig. 3: Histogram, detailed.



second and third. However, we could also imagine several books that share a higher number of trigrams, where the first ranked titles share ten trigrams, the next six etc. Thus, the connections between the publications can be ranked, and their relative strength can be measured. This enables us to make specific selections, based on these parameters. The next section describes some examples.

4. Finding Connected Titles

Using the data about the relative strength of the connections, it is possible to select publications based on several options. The first example consists of the titles connected to a single book. This could be used for recommender systems, showing a few closely connected titles to a book. After that, we will explore other possibilities, based on groups of publications.

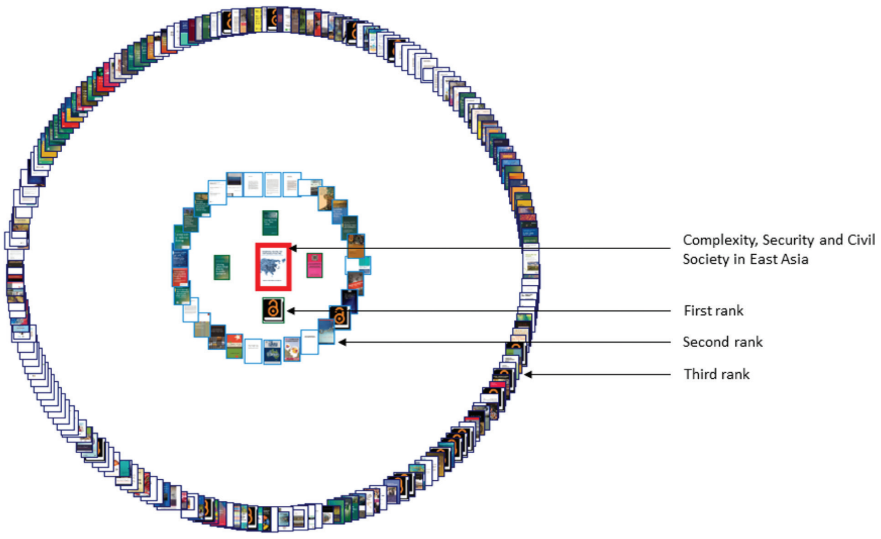
4.1. Single Book

This example is based on the book "Complexity, Security and Civil Society in East Asia" (Hayes & Yi, 2015), which discusses complex global problems such

as urban insecurity, energy, and climate change. It shares three trigrams with four titles, one of them is the book "Loss and Damage from Climate Change" (Mechler et al., 2019). These four titles are part of the first rank. Moving on to the second rank, there are 29 tiles. One of those titles is "Louisiana's response to extreme weather" (Laska, 2020). Furthermore, it shares one trigram with 217 titles, among them the book "Sustainable rice straw management" (Gummert et al., 2020); here the connection with insecurity and climate change seems weaker. However, the trigram both books share is "greenhouse gas emissions".

Figure 4 displays the three ranks:

Fig. 4: Book with related titles, ranked.



When looking at the connection between this book and the closely related books, the first question is what trigrams they share. These are listed in Table 2, *Shared trigrams*. The common theme connecting these books is quite clear: global warming and its effects.

However, the book "Complexity, Security and Civil Society in East Asia" does not only focus on climate change, and the trigrams reflect that. The most

Table 2: Shared trigrams.

Trigram	Title		
	Loss and damage from climate change	Louisiana's response to extreme weather	Sustainable rice straw management
greenhouse gas emissions	X	-	X
climate change adaptation	X	X	-
sea level rise	X	X	-

common trigrams are "civil society organizations" (occurring 78 times); "rok foreign policy" (occurring 57 times) and "world economic forum" (occurring 43 times). The first 'shared' trigram is "greenhouse gas emissions", which is mentioned 29 times. The term "climate change adaption" is mentioned 20 times – the almost identical trigram "climate change mitigation" was counted 17 times. Lastly, "sea level rise" could be found 14 times.

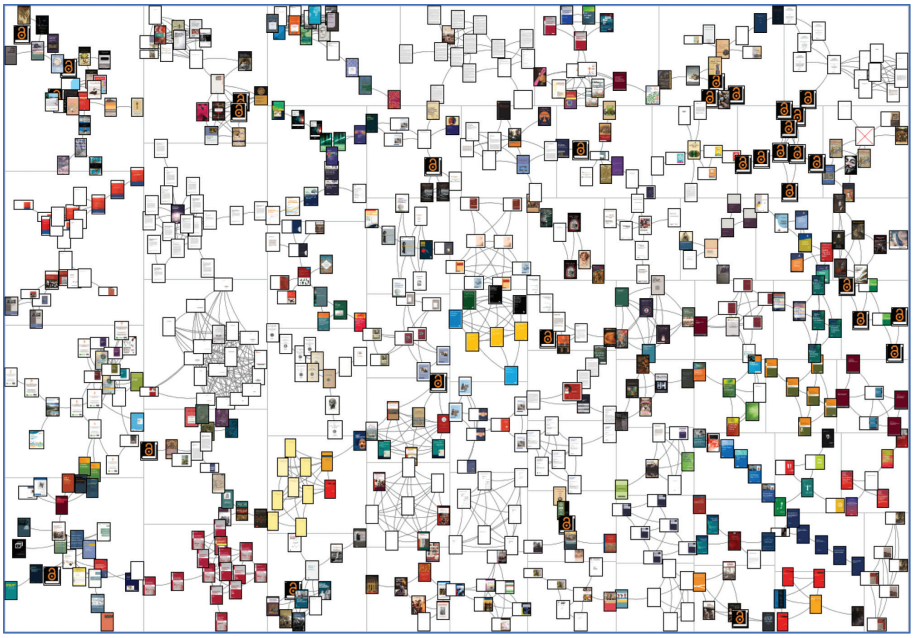
It is also interesting to look at which trigrams could not be linked. Several of them are related to policy making, which became clear from the top three trigrams and several mentions of the Nautilus Institute for Security and Sustainability, a public policy think-tank. Furthermore, nuclear energy and energy security are also mentioned in several trigrams. The complete list can be found in section Appendix.

4.2. Groups

The previous section showed the titles related to one book. Another possibility is to examine groups of publications and their relations. What books are closely connected, and does their relative 'distance' display subtopics within a larger collection? Figure 5 shows a selection of books that share three trigrams or more. Each of the groups consists of closely connected books and chapters.

Randomly selecting titles based on the number of shared connections does not lead to very useful results. Starting with one book, it makes sense to search for related titles. In order to find more relevant results for groups,

Fig. 5: Groups of books sharing three or more trigrams.

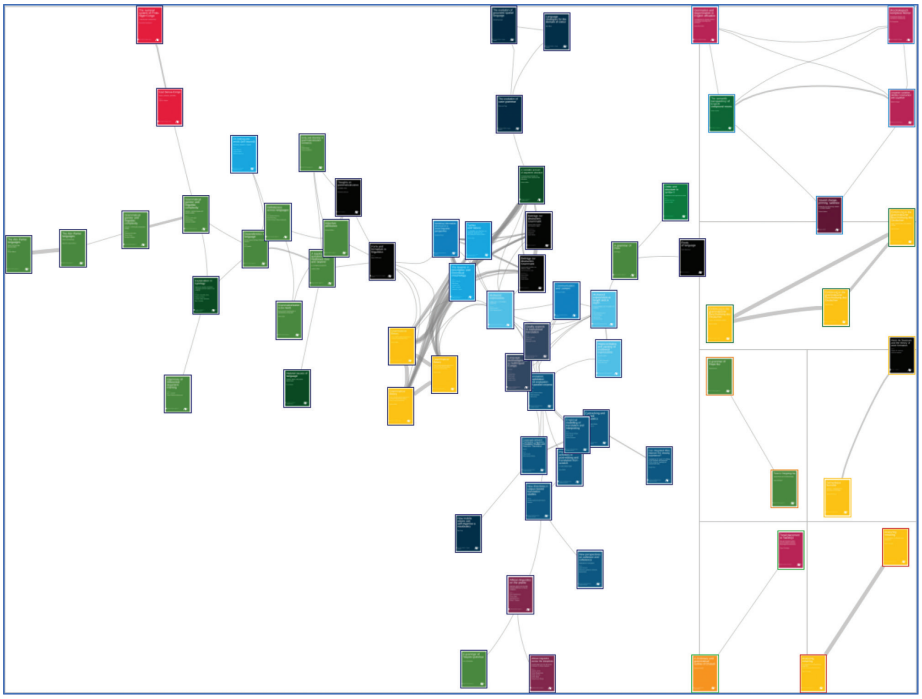


it is necessary to use additional metadata. In this case, the metadata of the OAPEN Library.

Using the metadata of the OAPEN Library enables us to search using several characteristics. In the next example, Figure 6 displays books published by Language Science Press. This publisher specialises in linguistics and all titles are part of a series; the colour of the cover denotes a series which helps to visualise the relations further. For instance, the green covers are part of the series "Studies in diversity linguistics", and the dark blue covers indicate the "Computational models of language evolution" series. Moreover, the thickness of the connecting line is an indication of the number of shared trigrams.

Instead of focussing on a single publisher, we could also look at the open access titles that received financial support from the same funder. If the funder has an underlying policy regarding the titles – see for instance Rieck (2019) – is that reflected in the publications? Figure 7 displays books funded by the Austrian Science Fund (FWF). Here, several smaller groups of closely connected books are noticeable.

Fig. 6: Connected books, published by Language Science Press.



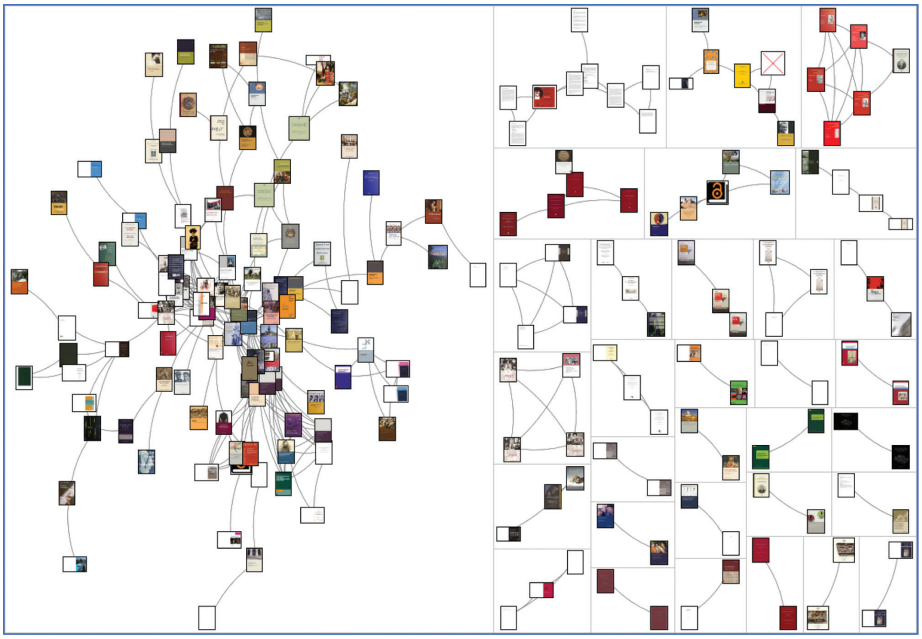
Furthermore, the two titles in the bottom right are translations: "Revolution and transition : Cultural policy in Bulgaria, 1989–2012" (Alexandrov, 2017a) and "Wende und Übergang : Die Kulturpolitik Bulgariens, 1989–2012" (Alexandrov, 2017b). The algorithm is capable of connecting books across languages. More on translations in the next section.

The graphics in this section were created using NodeXL (Smith et al., 2010). The data set and the algorithm in the R language is available at https://doi.org/10.17026/dans-xbm-qr5e.

4.3. Finding Translations

The connection between the two translated books in the set of FWF funded titles is not a coincidence. Within the data set, at least 15 "translated couples"

Fig. 7: Connected books, funded by FWF.



could be found. This might seem counterintuitive: the algorithm is based on finding exact trigrams, and one would expect translations to use different words to describe the same concepts. However, the analysis of several sets of translated books that share nine or more trigrams shows they often share English language terms, such as "adaptive cruise control" (Maurer et al., 2015, 2016); "labour force survey" (Holtslag et al., 2012, 2013) or "deep packet inspection" (Sprenger, 2015a, 2015b). Nevertheless, the shared terms do not have to be restricted to English, such as "graf leo thun" (Aichner & Mazohl, 2017a, 2017b). Additionally, web addresses also function as a language agnostic identifier. See for instance "http://www.siebenbuerger.de zeitung" (Hermanik, 2016a, 2016b) or "http://www.minfin.bg bg" (Alexandrov, 2017a, 2017b).

4.4. Use Cases

The previous sections described some of the possible applications of the trigram algorithm, based on a single books or groups of titles. What are possible

use cases for the stakeholders involved? The first use case is based on the connections surrounding a single title. As discussed in the introduction, this can be used to create a recommender system. For each title, the recommender system might display titles ranked first to third. The selection could also be refined by the number of titles: in the example of section 4.1, the number of third ranked titles linked to the book is 217, which is possibly too much for a single recommendation.

Creating benchmarks for publishers would be another use case. Here, the goal is comparing usage data of a set of comparable titles to a publication. By selecting all connected titles and collecting usage data it is possible to establish the average usage for this particular publication. This can be used as benchmark. Again, the number of titles to include can be varied by selecting only higher ranked titles.

Libraries might be interested in creating a collection of connected titles. Using the metadata such as keywords or classification creates a core set of titles, which can be expanded by selecting connected titles. Once more, the differences in ranking help to determine the extensiveness of the collection. A similar approach could also be used by researchers, looking for related titles to be used for citation or usage analysis.

5. Conclusion

Recommender systems based on personal data are successful but are not a viable option for those who want to protect the privacy of their users. Deploying a ngrams based algorithm is a good alternative for open access books, as it uses the contents of the publications. The algorithm quantifies the connections between the titles, which makes it easy to select a level of connectivity. The results can be used in several scenarios: recommendations for a single title or creating collections based on several conditions.

The use of trigrams and the algorithm to find related titles does not have to be confined to the OAPEN Library. The same method can be applied to other collections of open access books or even open access journal articles. By combining the trigrams and searching for matching titles, the algorithm helps to find relevant titles across multiple collections, enhancing its effectiveness.

Acknowledgments

The author would like to thank the colleagues of the OAPEN Foundation and Professor Martin Paul Eve of Birkbeck, University of London for commenting on previous versions of this paper. The data of this paper is publicly available through the support of Data Archiving and Networked Services (DANS).

References

Aichner, C., & Mazohl, B. (2017a). *Die Thun-Hohenstein'sche Universitätsreformen 1849–1860: Konzeption – Umsetzung – Nachwirkungen*. Böhlau. https://library.oapen.org/handle/20.500.12657/31673

Aichner, C., & Mazohl, B. (2017b). *The Thun-Hohenstein University reforms 1849–1860: Conception – Implementation – Aftermath*. Böhlau. https://library.oapen.org/handle/20.500.12657/31171

Alexandrov, A. (2017a). *Revolution and transition: Cultural policy in Bulgaria, 1989-2012*. LIT Verlag GmbH & Co. KG. https://library.oapen.org/handle/20.500.12657/31422

Alexandrov, A. (2017b). *Wende und Übergang: Die Kulturpolitik Bulgariens, 1989-2012*. LIT Verlag GmbH & Co. KG. https://library.oapen.org/handle/20.500.12657/31423

American Library Association. (2014). *Privacy: An interpretation of the Library Bill of Rights*. http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy

Corrado, E. M. (2007). Privacy and Library 2.0: How do they conflict? *Ailing into the Future: Charting our destiny: Proceedings of the Thirteenth National Conference of the Association of College and Research Libraries*. https://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/national/baltimore/papers/330.pdf

Eve, M. P. (2019). Reading genre computationally. In M. P. Eve (Ed.), *Close Reading with Computers: Textual scholarship, computational formalism, and David Mitchell's Cloud Atlas* (pp. 61–95). Stanford University Press. https://doi.org/10.21627/9781503609372

Gummert, M., Hung, N. V., Chivenge, P., & Douthwaite, B. (2020). *Sustainable rice straw management*. Springer Nature. https://doi.org/10.1007/978-3-030-32373-8

Hayes, P., & Yi, K. (2015). *Complexity, Security and Ctivil Society in East Asia: Foreign Policies and the Korean Peninsula*. Open Book Publishers. https://www.openbookpublishers.com/product/326

Hermanik, K.-J. (2016a). *Deutsche und Ungarn im südöstlichen Europa: Identitäts- und Ethnomanagement*. Böhlau. https://library.oapen.org/handle/20.500.12657/31956

Hermanik, K.-J. (2016b). *Germans and Hungarians in Southeast Europe: Identity management and ethnomanagement*. Böhlau. https://library.oapen.org/handle/20.500.12657/29393

Holtslag, J. W., Kremer, M., & Schrijvers, E. (2012). *In betere banen*. WRR. https://doi.org/10.26530/OAPEN_440005

Holtslag, J. W., Kremer, M., & Schrijvers, E. (2013). *Making migration work: The future of labour migration in the European Union*. Amsterdam University Press. https://library.oapen.org/handle/20.500.12657/33887

International Federation of Library Associations and Institutions. (2016). *IFLA Code of ethics for librarians and other information workers (full version)*. http://www.ifla.org/publications/node/11092#privacy

Jaeger, P. T., McClure, C. R., Bertot, J. C., & Snead, J. T. (2004). The USA PATRIOT act, the foreign intelligence surveillance act, and information policy research in libraries: Issues, impacts, and questions for libraries and researchers. *The Library Quarterly*, *74*(2), 99–121. https://doi.org/10.1086/382843

Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., & Tang, Q. (2013). Privacy in recommender systems. In N. Ramzan., R. van Zwol., J.-S. Lee., K. Clover., & X.-S. Hua (Eds.), *Social media retrieval* (pp. 263–281). Springer. https://doi.org/10.1007/978-1-4471-4555-4_12

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, *3* (pp. 255–264). https://web.cs.dal.ca/ vlado/papers/pacling03.pdf

Laska, S. (2020). *Louisiana's response to extreme weather: A coastal state's adaptation challenges and successes*. Springer Nature. https://doi.org/10.1007/978-3-030-27205-0

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Maceli, M. G. (2018). Encouraging patron adoption of privacy-protection technologies: Challenges for public libraries. *IFLA Journal*, *44*(3), 195–202. https://doi.org/10.1177/0340035218773786

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2015). *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Springer Nature. https://doi.org/10.1007/978-3-662-45854-9

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2016). *Autonomous driving: Technical, legal and social aspects*. Springer Nature. https://doi.org/10.1007/978-3-662-48847-8

Mechler, R., Bouwer, L. M., Schinko, T., Surminski, S., & Linnerooth-Bayer, J. (2019). *Loss and damage from climate change: Concepts, methods and policy options*. Springer Nature. https://library.oapen.org/handle/20.500.12657/23027

Miao, Y., Kešelj, V., & Milios, E. (2005). Document clustering using character N-grams: A comparative evaluation with term-based and word-based clustering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 357–358). http://www.ezcodesample.com/SemanticSearchArt/downloads/CS-2005-23.pdf

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Pazzani, M., & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky., A. Kobsa., & W. Nejdl (Eds.), *The Adaptive Web* (pp. 325–341). Springer. https://doi.org/10.1007/978-3-540-72079-9_10

Project MUSE. (2021, June 21). *Project MUSE introduces AI-based links, powered by UNSILO, for related content*. https://about.muse.jhu.edu/news/unsilo-ai-based-links-on-muse/

R Core Team. (2020). *The R project for statistical computing.* The R Foundation. https://www.R-project.org/

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender systems handbook*. Springer. https://doi.org/10.1007/978-0-387-85820-3

Rieck, K. (2019). The FWF's Open Access Policy over the last 15 Years – Developments and Outlook. *Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare*, *72*(2). https://doi.org/10.31263/voebm.v72i2.2837

Rohini, U., & Ambati, V. (2007). Extracting Keyphrases from books using language modeling approaches. In *Proceedings of the 3rd International Conference on Universal Digital Library*. ulib.isri.cmu.edu/conference/2007/Rohini.pdf

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on electronic commerce* (pp. 158–166). ACM. https://doi.org/10.1145/336992.337035

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, *1*(3). 37. https://doi.org/10.21105/joss.00037

Silge, J., & Robinson, D. (2017). *Text Mining with R: A tidy approach* (1st ed.). O'Reilly Media. https://www.tidytextmining.com/

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE Internet Computing*, *21*(3), 12–18. https://doi.org/10.1109/MIC.2017.72

Smith, M., Ceni, A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Lescovec, J., & Dunne, C. (2010). *NodeXL: a free and open network overview, discovery and exploration add-in for Excel*. Social Media Research Foundation. http://www.smrfoundation.org

Snijder, R. (2019). Patterns of information—Clustering books and readers in open access libraries. In *The deliverance of open access books: Examining usage and dissemination* (pp. 83–103). Amsterdam University Press. https://doi.org/10.26530/OAPEN_1004809

Snijder, R. (2021). *OK Computer, what are these books about? – An experiment in large-scale classification of open access books*. Manuscript submitted for publication.

Souza, R. R., & Raghavan, K. S. (2014). *Extraction of keywords from texts: An exploratory study using noun phrases*. https://hdl.handle.net/10438/28306

Sprenger, F. (2015a). *Politik der Mikroentscheidungen: Edward Snowden, Netzneutralität und die Architekturen des Internets*. meson press. https://library.oapen.org/handle/20.500.12657/37577

Sprenger, F. (2015b). *The Politics of Micro-Decisions: Edward Snowden, net neutrality, and the architectures of the Internet*. meson press. https://library.oapen.org/handle/20.500.12657/37575

Witt, S. (2017). The evolution of privacy within the American Library Association, 1906–2002. *Library Trends*, *65*(4), 639–657. https://doi.org/10.1353/lib.2017.0022

Notes

[1] Recently, Project MUSE announced a recommender system based on artificial intelligence (Project MUSE, 2021).

[2] OAPEN Foundation. *OAPEN Library*. https://www.oapen.org.

[3] See https://scholar.google.com/scholar?hl=en&as_sdt=0,5&q=%22Google+books+ngram%22.

Appendix

Trigrams of the book "Complexity, Security and Civil Society in East Asia".

Trigram	Amount
civil society organizations	78
rok foreign policy	57
world economic forum	43
civil society networks	35
greenhouse gas emissions	29
berkeley nautilus institute	24
east asia institute	23
jeju peace forum	23
north korean nuclear	21
climate change adaptation	20
doi http dx.doi	20
green economy policies	19
climate change mitigation	17
energy security policies	17
seoul nautilus institute	17
united nations development	17
nautilus institute 2010	16
2011 doi http	15
geneva world economic	15
napsnet special report	15
nuclear fuel cycle	15
nuclear power plants	15
security seoul nautilus	15
york united nations	15
energy security seoul	14
energy supply security	14
sea level rise	14
yonhap news agency	14
2008 doi http	13
east asia green	13