Forward-looking analysis based on grants data and machine learning based research classifications as an analytical tool

Christian Herzog, Aaron Sorensen, ÜberResearch GmbH

Michael Taylor, Digital Science Ltd

The challenge

In the past, the outcomes of funded research projects have been the primary source of data used to assess the impact and outcome of science policy. The reason behind this choice has been one of convenience. Firstly, the data and documents in question are available and easy to use for analysis, and secondly, there exist widely-accepted metrics for understanding the volume of output, as well as some means of understanding the quality and potential impact by analysis of the citations accrued by scientific publications.

Systematically analysing the output of funded projects will remain an important tool in the analyst's armory, but it is important that the limitations of this post-hoc analysis are understood and challenged.

For example:

- **The link between publications as the output and the source of funding is provided by the author or the authors is not straightforward.** Since multiple authors might be funded by several funders and a publication might be supported by multiple projects there is not an easy one-to-one relationship between funders, researchers and published output, let alone downstream impact. The obvious challenge is how to develop sophisticated methodologies to understand these relationships, whether to single publications or multiple projects.

- **Focusing on publications introduces an inherent time lag of 3-8 years into any given analysis.** After being awarded a given grant, the investigator starts to do the research, preliminary results may be presented at conferences before finally the results are published typically as a journal article. The publishing process - which may involve many stages of peer-review and revision - can be an eighteen-month process. After an article is published, it then takes at least another 2 years to allow the research community to begin showing their approval or indifference for the published work by citing (or failing to cite) the article, a further delay may be incurred by the time taken by

the major indexing systems to calculate relevant metrics.

- **The relationship between citation and socio-economic impact is not certain.** Increasingly funders of research, such as the European Commission, are expecting researchers to demonstrate the impact of their outputs in a wider context than purely scholarly. Whether this broader impact is principally cultural, social, economic or industrial, there is a dearth of evidence to suggest that citation-based analysis can provide detailed insights into these forms of broader impact.

  Furthermore, the dependence on a database of publications and citations means that efforts on behalf of researchers to promote their research outside the research community may be neglected. As policy-makers are often influenced by social discourse in the mass media and on social networks, we potentially lose sight of the full picture of the research lifecycle.

These disconnects - between the fast changing and fast growing field of scientific research and the tools that we have to measure and understand its current state - present a serious challenge to people who are seeking to influence or direct the direction of research.

Understanding current trends in global grants gives a window into future research outputs

In an ideal world, one would be able to complement the analysis of the outputs with a view on the 'inputs' – therefore allowing:

1. an analyst to see in which areas of science - and in which specific projects - scientific resources are being allocated today, and therefore
2. being able to get a view on the research which will be carried out in the coming years.

Providing this forward view would provide data to analysts to present a more balanced view for science policy decisions: not only what has been achieved over the last decade, but also what is going to be done in the coming years.

The solution is far from trivial. A system would:

1. require a database of funded projects from as many funders as possible and
2. the ability to analyse the proposed science in depth, which, in turn,
3. would require techniques such as natural language processing to parse project descriptions.

In absence of a publicly-available, funded-projects database, the team from ÜberResearch, together with funders as development partners, began in 2013 to build up just such a project database.

As of March 2016, this global research-project repository has grown to more than 2.1 million projects covering more than $1 trillion of funded research from more than 200 funders globally. Out of the total, $282 billion of funding is for projects which are active in 2016 and in the coming years.

This allows for a detailed, forward-looking analysis of almost any scientific discipline. Using these data, a science policy-maker may:

1. Discover what research is currently funded and underway,
2. What research will be carried out in the future, and
3. Can drill-down through the aggregate values to display researcher- or institution-level activity in any given field.

Having built the database and established the relationships, several analyses may be undertaken:

- The data provides for a view of the future work and outputs,
- It allows one to analyse the funding decisions and investments other bodies have made,
- A planned program of Funder A may be compared against the funded projects of other global organizations.

The challenges of funder-specific research classification systems

Most funders use bespoke research classification systems which are mostly applied manually either by the applicants themselves or by subject matter experts, assigning the relevant codes. This has several implications:

- Only a limited number of research classification systems can be used due to the efforts and costs involved (typically just a single classification system is used by any given funder) - meaning that there are either compromises in granularity and accuracy, or a high barrier to effective encoding.
- Research classifications are only applied to small document sets (e.g. only one's own portfolio, not other funders' portfolios or publications) due to the efforts and costs involved - implying that any analysis will be partial.
- Inconsistent coding: the assignment decisions of human coders (i.e. inter-rater reliability) can vary immensely – making a comparison and analysis difficult.

In the absence of a global system of classifying research subjects - and, more challengingly, a method for maintaining, improving and expanding a system, new techniques have to be found to support cross-portfolio, international funding decisions and applications.

Fortunately, the technology of machine-learning allows for the development of cheaper, faster classification that learns from the experience of human subject experts and allows for the consistent coding of large document sets with much lower unit costs.

Research classification based on machine learning – an approach for portfolio analysis

With the support of many funders, ÜberResearch has an analytical tool called "Dimensions" on top of the grant database. Dimensions allows a policy analyst to search and quickly understand the funding landscape in any area of science. Inconsistent metadata, the lack of a common taxonomy mean that generic search capabilities do not allow for a systematic analysis of the portfolio of a given funder, or for a profound comparison of the activities of two research funding organizations.

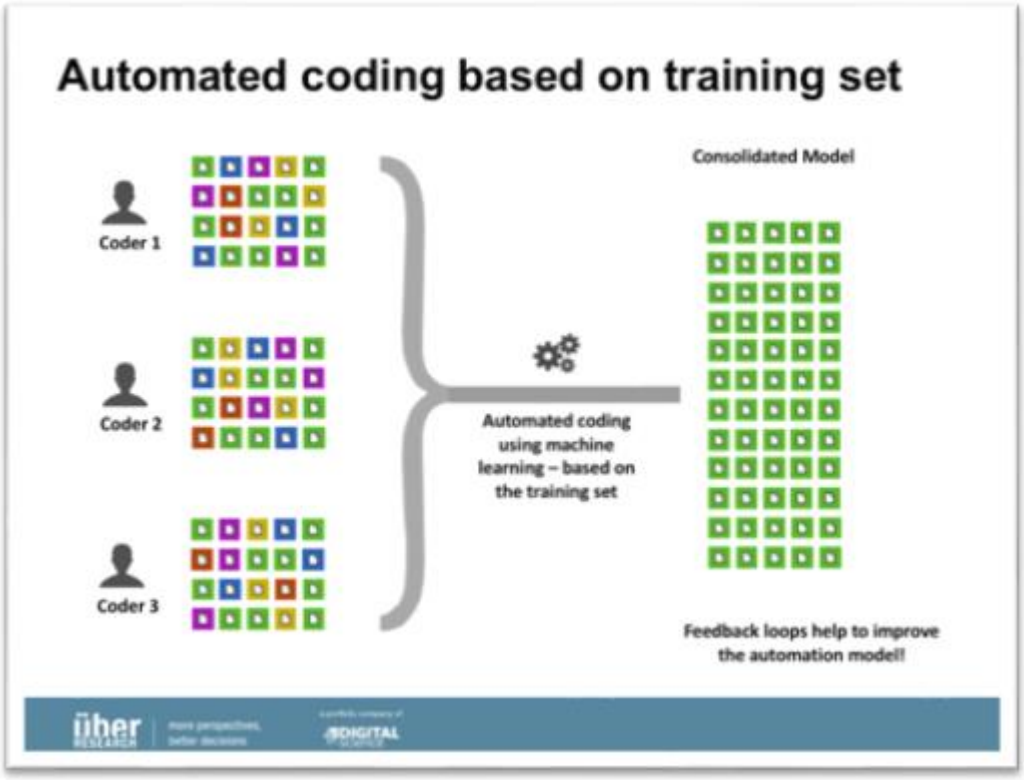

Fig 1: using human coded training sets as a basis for the machine-learning-based model

Working with funders, ÜberResearch analysed their various classification systems and applied machine learning techniques. This technology takes human classified documents and analyses the content for correlations and other statistical relations to develop rules and models based on the human decisions. The system then goes through a process of applied, refining and developing new models to improve its accuracy before going into production. More specifically, our classification pipeline is based on a bag-of-words approach with term-frequency-inverse-document-frequency (TF-IDF) weighting scheme (e.g., Salton and Buckley, 1988). The final step of the pipeline employs an ensemble of linear support vector machines (SVMs; Cortes and Vapnik, 1995). SVM was chosen as the most proven machine learning classification algorithm for textual data (Joachims, 1998; Yang and Liu, 1999).

For the two-stage budget-division setup we implement a gates-and-experts methodology (Shen et al., 2011) where the problem is decomposed into two models, the first doing coarse-grained predictions (budgets) and the second conducting fine-grained predictions. The advantage of using this method over a single-model classifier that learns all class combinations at once is that both stages have fewer classes to deal with and the second-stage model can be learned only on a specific subset of the training data which improves the quality of generated discriminative features.

The ÜberResearch system has implemented various classification systems as machine learning based models, which are now applied simultaneously and instantly to all projects in the database (regardless of funder) to be leveraged by all end users (regardless of employer). Some examples are listed in Table 1.

• Fields of Research from Australia and New Zealand (FOR): covering all areas of science[1].
• Health Research Classification System from the UK[2], a system used by the Medical Research Council and other biomedical funders to report on activities. The Health Research Classification System (HRCS) is a two dimensional framework - codes from both HRCS dimensions are applied when classifying. One dimension, the Health Categories, is used to classify the type of health or disease being studied. There are 21 categories encompassing all diseases, conditions and areas of health. The other dimension and, the Research Activity Codes, classifies the type of research activity being undertaken (from basic to applied).
• Common Scientific Outline from the International Cancer Research Partnership[3]: (Awards on the International Cancer Research Partnership (ICRP) database are coded using a common language — the Common Scientific Outline or 'CSO', a classification system organized into six broad areas of scientific interest in cancer research. The CSO is complemented by a standard cancer type coding scheme. Together, these tools lay a framework to improve coordination among research organizations, making it possible to compare and contrast the research portfolios of public, non-profit, and governmental research agencies.
• RCDC (Research, Condition, and Disease Categorization)[4] is a computerized reporting process the National Institutes of Health (NIH) is using at the end of each fiscal year to categorize its funding in medical research beginning with fiscal year 2008 RCDC (https://report.nih.gov/rcdc/) reports NIH funding in 233 research, condition, and disease categories.

Table 1: Examples of classification systems implemented by ÜberResearch's machine learning system

This implication of this work is that Funder A's grants and outputs - which are natively classified only in their bespoke classification model - may be re-classified and analysed by Funder B using Dimensions - using their own classification model - and vice versa.

The machine learning technology and the global grants database are combined to make a unique analytical tool for scholarly funding research.

Combining the machine learning models with the analytical side of Dimensions

ÜberResearch has developed a process and pipelines which allow us the ability to generate the machine learning models with the involvement and feedback of the subject matter experts in a standardized way. This allows funders to develop and use their own classification systems or adopt existing ones at marginal costs. The models are integrated into the analytical tool, Dimensions, allowing for immediate analyses to be undertaken.

[1] see also http://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA257418000446

[2] http://www.hrcsonline.net/

[3] https://www.icrpartnership.org/CSO.cfm

[4] https://report.nih.gov/rcdc/

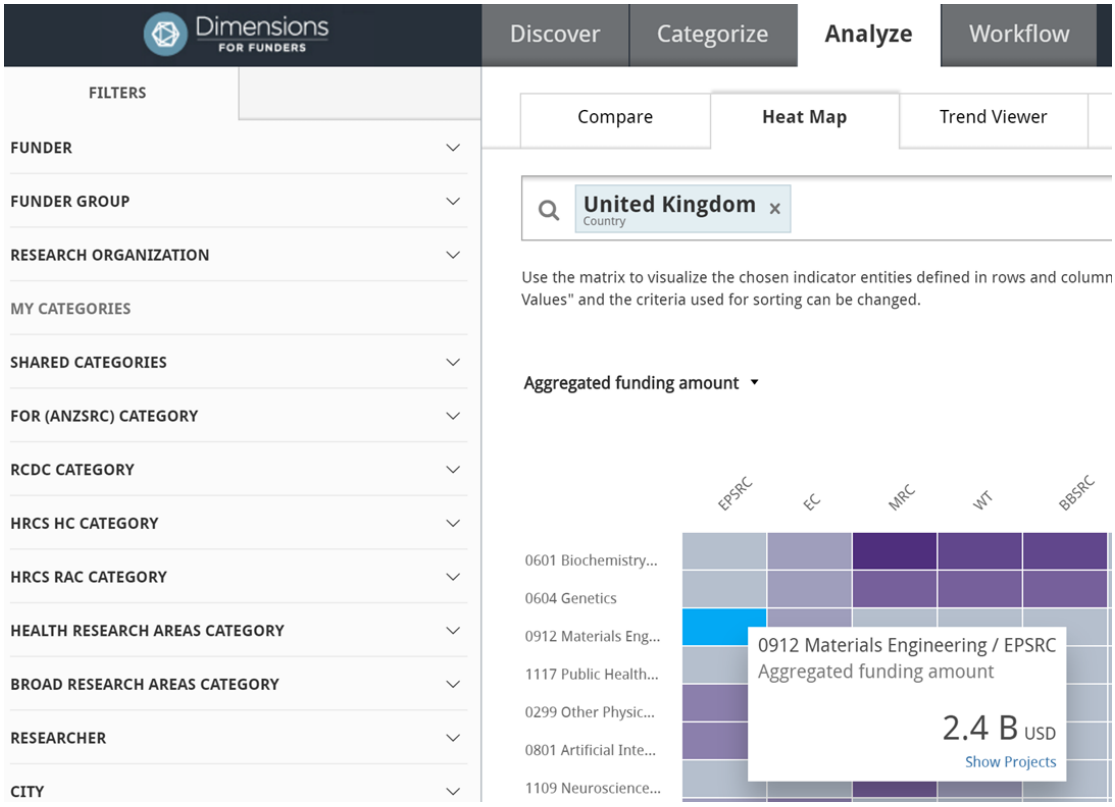

Fig.2 View on funders funding research organizations in the UK per FOR categories

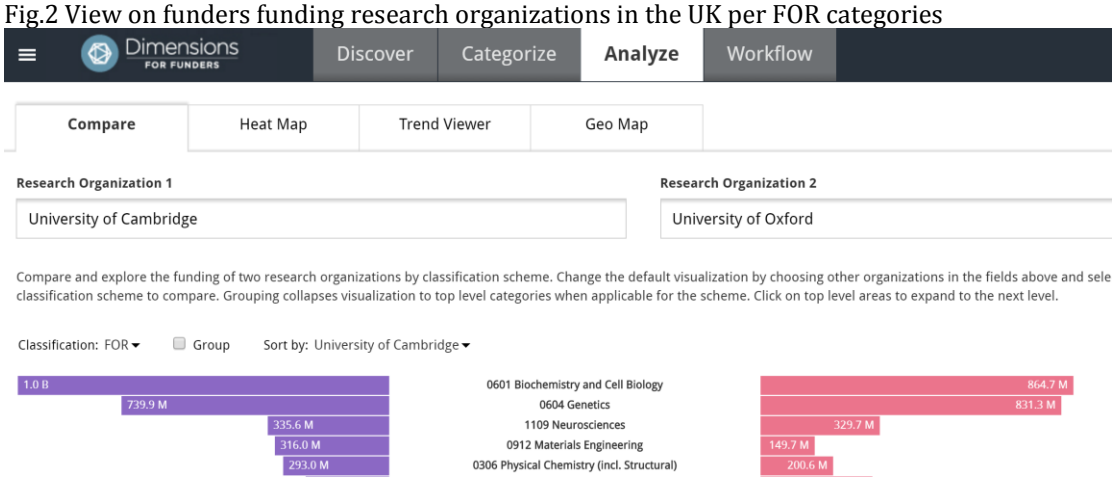

Fig. 3: FOR-based comparison of the portfolios of the University of Oxford and University of Cambridge

Looking at the two screenshots above, one could imagine a scenario in which someone involved in making UK science policy is looking at the landscape of funded research in the UK from the perspective of funders. For example, one might ask which British funder is currently investing the most in the area of Materials Engineering and from the visualization in Figure 1 quickly come to the conclusion that the Engineering and Physical Sciences Research Council (EPSRC) is leading the way in this area. One might then wish to turn his attention to the university side of the analysis. When examining research in the UK, comparisons between Cambridge and Oxford abound, so in Figure 3, we extend our exploratory scenario in which our analyst decides to see whether Cambridge or Oxford is receiving significantly more funding for Materials-Engineering research than the other. From Figure 3, we can see that in the last decade, Cambridge has

received approximately twice the material-engineering grant support when compared to arch-rival Oxford.

Applying the machine-learning based models across publications, grants, and other documents to create a consistent portfolio view across inputs and outputs

With the machine-learning-based models, it is possible to apply the same categorization approach to different document sets, for example to grant descriptions as inputs and to publications as an output. The approach creates comparable data sets where natural language processing and machine learning are used to tap into the 'substance' of the research – allowing for immediate and deep insights.

The Challenge of Understanding Broader Impact and the Increasingly Diverse and Dynamic Scientific Network

Obtaining data, creating relationships and automatic classification of topics are essential steps in developing analyses of current research and future output. However, as data and technology improves - and the community becomes more adept in using future-looking technology, we may expect to see more sophisticated analytical tools emerge.

There is considerable work to be done to obtain a fuller picture of the effect that grant-policy and funding-decisions has on scientific discovery and broader impact, and two fields may provide impetus towards those future developments. By the same token, these two fields can be used to shed light on conversations and documents that may be used to shape funding decisions and grant policies.

Altmetrics, grey literature and broader impact

The idea of 'measuring by proxy' is not an uncommon one to research evaluation professionals and bibliometricians. However inadvisable it may be, for decades attempting to measure the quality of a researcher's output by means of the Journal Impact Factor of the journals in which they publish has been a key component for many people. Similarly, attempting to measure the broader impact of research in society by means of citation analysis is challenging. While there have been attempts - for example, analysing the number of non-academic co-authors on a research paper - the data is necessarily limited to the venue in which scholarly publishing occurs. Citation, publishing and co-authorship analyses are entirely limited to a closed network.

The last six years have seen an enormous increase in the interest being paid to metrics that are derived from sources outside this network. A recent measure of this increased attention has been the European Commission's establishment of an Altmetric Experts Group[5]. This field has seen a burgeoning number of conferences, papers and research activities to develop theory and observations that may be useful in expanding the analysis of grant-decisions and funding policies outside the closed network.

Although 'altmetrics' is frequently discussed as if it is a homogenous phenomenon, it covers a wide variety of data types, potentially describing a number of behaviours:

[5] Expert Group on Altmetrics, http://ec.europa.eu/research/openscience/index.cfm?pg=altmetrics_eg

- Sharing and saving on platforms such as Papers, Mendeley and Citeulike is typically undertaken by people engaged in research. Evidence (Mohammadi et al 2015] indicates that although users are working with research outputs, their use may not be reflected in citation rates. As well as academic researchers, other users include industry researchers, non-academic specialists and students. Data is generally limited to activity counts.
- Sharing links on general purpose social networks, such as Twitter or Facebook may provide quantitative data (number of Tweets, re-tweets etc) and qualitative data (the content and the analysis of the content).
- A selection of the so-called grey literature is included in Altmetric data, where those documents link to research output. This might include patent data, academic blogs and policy documents. A highly relevant data source might be citations and links made to academic papers within funding documentations.
- Coverage of and links to academic papers in the mass media is naturally low, however may be considered to be potentially highly impactful.

In each case, the data may be examined for evidence of impact, and by connection to research output, to funding decisions and grant policy. Although the amount of data being collected and reported is growing every year, it is partial and somewhat skewed towards certain disciplines. And, of course, it must be understood that research in different fields does not have a uniform potential for immediate broader impact.

There are two further issues with understanding the impact of research in governmental and non-governmental policy papers. Firstly, and perhaps understandably, the authors of these documents rarely cite or link to academic papers, thus making it extremely difficult to connect research output to policy, thereby affecting our ability to report on the influence of upstream funding decisions on downstream governmental and non-governmental policy. Where such links or citations are made, they very often provide inconsistent metadata, making the resolution of the citation highly challenging.

A second, and perhaps more critical attribute of policy papers is their impermanence. Documents are frequently difficult to find in repositories, and often disappear. This makes establishing a record of influence a great challenge for those of us involved in mapping broader impact.

Despite these difficulties, altmetrics and allied developments do give us the possibility that we might be able to expand our view of the impact of scientific research into arenas that have a broader impact than the purely scholarly.

When focussing on the question of grey literature and using it to understand the future of research, consideration must be given to the location of conversations about the needs of society and the potential direction of research. It is likely that as well as using altmetrics to measure the downstream impact of research in a broader context, it will be possible to model the conversations that influence upstream funding policy.

The question is whether technology exists to understand and model these discourses, and whether the discourses themselves are held in a reliably open and accessible venue. Indeed, if open science is to become truly effective, then the documents and decisions that govern policy and the application of research must become more open themselves.

A computable network of literature: towards understanding the dynamics of funding and broader impact through network analysis

The science of bibliometrics has traditionally taken a linear approach to understanding impact. This is not for want of imagination amongst that field, rather the complexities of understanding a

world that consists of millions of documents and billions of connections, and one that grows at an ever-increasing rate.

In the linear world, a researcher usually applies for a grant, receives funding (or is otherwise employed), undertakes research, analyses data, publishes the findings in a journal, receives citations and returns back to the start of the process. The process is slow and complicated, and has a number of dependency stages which can make the process even more complex, and harder to analyse to obtain a forward-looking picture on which to basis grant decisions.

If altmetrics hints at a future where we can gather more data - and gather it faster - and detect influential conversations in venues that fall outside the closed scholarly network, then graph mathematics and network theory provide us with an opportunity to understand a field in a more holistic approach. It is hoped that by using big data approaches, that combine with sociological theories of knowledge diffusion, a blended approach that builds on the strengths of bibliometric expertise may be developed.

If theories such as Latour's Actor-Network Theory (LaTour 2005) suggests tantalizing possibility that scholarly literature, citation, social network content, policy papers and grant documents themselves might be treated as dynamic elements in an open model of scholarly knowledge, then approaches such as graph mathematics gives us the possibility that we may compute and quantify the qualities inherent in such networks.

The fact that networks of documents have particular mathematical properties has long been understood: probably the most noteworthy implementation is Google's PageRank algorithm[6]. PageRank computes authority, on the basis of the authority links that are made to web pages. Elegantly, it was originally inspired on the work of citation analysis.

However, PageRank - and authority - are not the only values that may be computed for a network. Mathematical techniques exist to quantify such values as proximity, centrality and direction; to automatically identify dynamic clusters of documents, and to compute their direction.

If a network is constructed that consists of scholarly and related documentations, and if connections are found, then it may become possible to analyse future trends of research and the effects of grant decisions and funding policies at great distance, and with increasing nuance.

In a full network view of scientific impact, we must move from a linear view of research and funding and towards a more complete - and computable - model.

Conclusion

Through our global, development-partner relationships with funders, we have come to understand the importance of integrating research-project descriptions with research outputs such as publications to allow for forward-looking science-policy analyses.

By way of example, imagine a novel grant being funded that goes on to produce a breakthrough finding which unfortunately goes unnoticed largely because it published in a low-Journal-Impact-Factor journal. Continuing with this scenario, the paper is subsequently picked up by an influential science blogger whose blog post inspires a policy-document writer to cite the paper in a policy document. The citation by the policy document increases the paper's Altmetric donut score, thus drawing the attention of a program officer at a key funding agency. The program officer then uses grant categorization to uncover the worldwide scarcity of funding for this niche

[6] https://en.wikipedia.org/wiki/PageRank

area of science and subsequently decides to issue a Request-for-Applications asking for a whole batch of grant submissions along the same lines. This time around, the authors of the papers generated by the grants in question have a much easier time getting their papers accepted by more prestigious journals.

Through this approach, we believe that scientometric analyses can begin to serve as inputs for "mid-course corrections" of recently enacted policy as opposed to being used for "post-mortem" evaluations of long-standing policies. As science becomes more open, and policies - as well as outputs and data - become more accessible, then we believe that future analyses will become more nuanced and accurate.

References

Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks. Mach. Learn.*, 20(3), 273–297.
Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning (pp. 137–142). London, UK, UK: Springer-Verlag.

LaTour, B., *Reassembling the Social*, 2005, OUP

Mohammadi E, Thelwall M; Kousha K (2015) *Can Mendeley bookmarks reflect readership? A survey of user motivations,* JASIST, http://onlinelibrary.wiley.com/doi/10.1002/asi.23477/full

Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval.* In Information Processing and Management (pp. 513–523).

Shen, D., Ave, E. H., Jose, S., & Sundaresan, N. (2011). *Item Categorization in the e-Commerce Domain Categories and Subject Descriptors.* CIKM, (88), 1921–1924.

Yang, Y., & Liu, X. (1999). *A Re-examination of Text Categorization Methods*. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 42–49). New York, NY, USA: ACM.