The Story of SAGE Recommends

Alan Maloney
Senior Product Analyst, Taxonomy & Semantic Technology

As a publisher of academic content that ranges from journal articles to numerical data to educational videos, covering subjects from theology to robotics, SAGE Publishing has always been interested in helping users to discover scholarly resources and connect concepts in new and innovative ways. In our 2016 discoverability white paper, *Expecting the Unexpected* (https://us.sagepub.com/sites/default/files/SerrDiscovery.pdf), we share research findings into serendipitous information seeking and explore potential solutions from publishers and librarians that support unplanned instances of scholarly discovery, such as content- and context-driven recommendations. In our research, we found that students and faculty are inclined to seek out content that they are comfortable with, for example, books and journals, without necessarily being aware of where to find other potentially useful sources of information—despite the fact that they state they would make use of a diverse range of formats (see Figure 1).

SAGE's new discovery tool, SAGE Recommends, was released in late 2015 to link together different types of content from across all SAGE products to expose users to novel, unexpected information based on the core principles of relevance and interestingness that our users have established as priorities. When relevance and interestingness are the criteria, we have found that there is no substitute for a content-based approach, especially as content is arguably the thing that a publisher should do best.

There are many ways of interpreting "content-based approach" or even "content." By "content," we mean an item that a user is currently engaging with, as distinct from all content that the user may have engaged with at some point in time. The latter poses several challenges: As well as user reservations with being tracked, there is also a danger that taking too much information into account leads to an overwhelmingly broad stream of recommended materials. YouTube is a good example: when logged in, users are presented with a huge variety of videos based on all of their previous viewing history, but when viewing a specific video, relevance is restricted to

Suggested Citation: Maloney, A. (2016). The story of SAGE Recommends (White paper). Thousand Oaks, CA: SAGE Publishing. doi: 10.4135/wp160129.2. Retrieved from https://us.sagepub.com/sites/default/files/SAGERecStory.pdf

what users are currently engaged with. We decided to focus on the latter use case, as our users tend to focus on one specific research question or information need at a time, rather than openly browse for entertainment. SAGE Recommends focuses on the content users are currently engaging with and aims to support contextualized, “just-in-time discovery,” rather than distracting users with their previous (and possibly fulfilled) information needs. We take a “content-based approach” to recommendation to mean various methods of statistical comparison between documents, of which perhaps the most common is to use natural language processing to transform a document into concepts (*vectors*) scored by how relatively frequently they appear (*tf-idf*) in comparison to a reference corpus. By converting a text document into vectors and numbers in this way, the statistical similarity between documents can be calculated and used as the basis for a content-based recommendation engine, as indeed many information providers do. This can be characterized as a *string-based* approach to document similarity (Gomaa & Fahmy, 2013).

String-based approaches to recommendation tend to work well for traditional academic content, which is why many publishers implement some form of string-based recommendation engine on their platforms. The main limitation of the approach is that simply converting a document into strings encodes no information about the meaning of those strings and consequentially whether any individual concept is related to another or, indeed, the same as another. Consider two authors writing about the decriminalization of a specific drug, one using the string “cannabis” and another using “marijuana.” Those two strings would not improve the similarity score between the two documents, despite them being, in essence, the same concept. This effect is usually mitigated in academic full-text content by the presence of other co-occurring strings; for example, although the authors use different terms for the same concept, they may both use strings like “decriminalization,” “possession,” or “medicinal,” which when used together give enough of a basis for the documents to be considered similar

Figure 1 Types of content that are important to students and faculty members

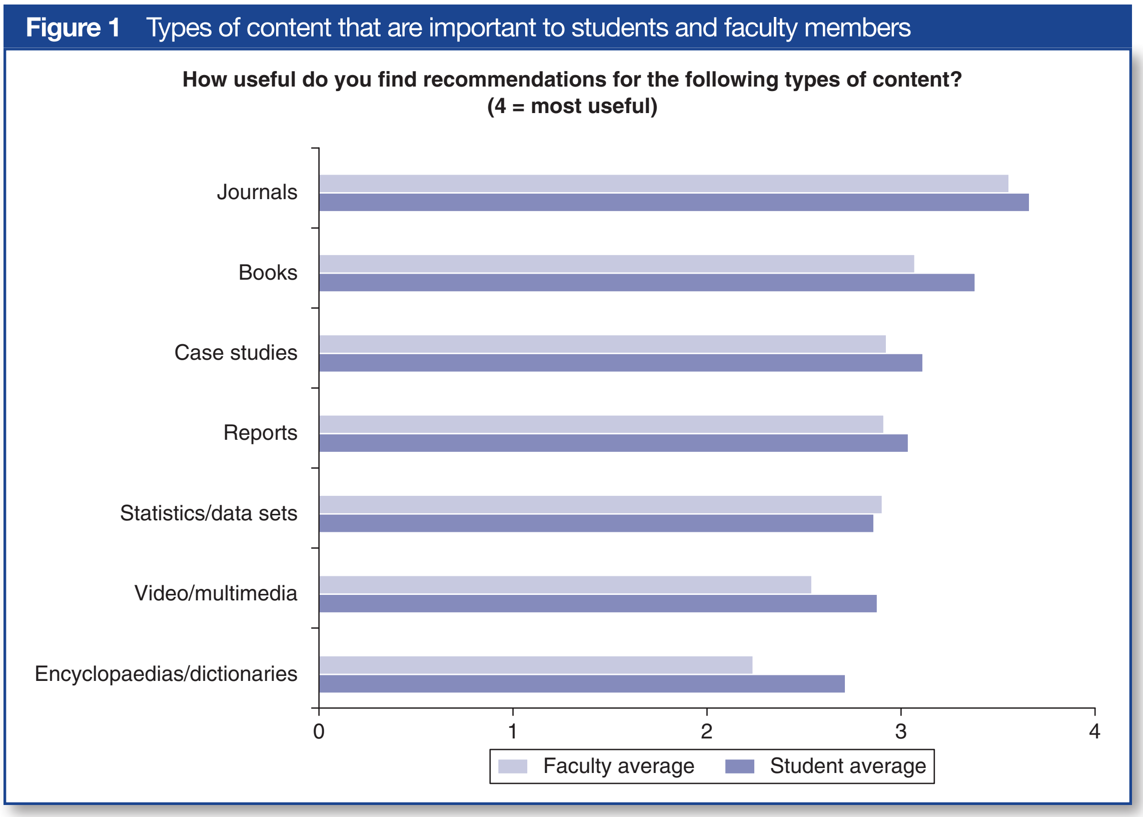
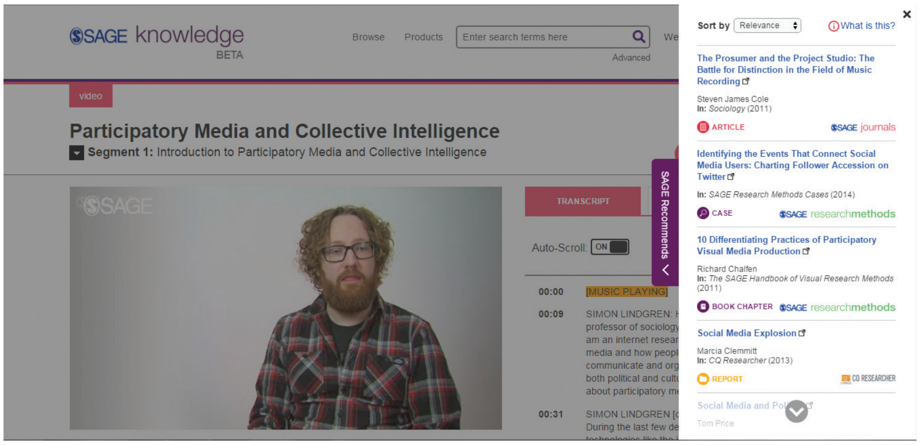


Figure 2 SAGE Recommends content based on a video



Participatory Media and Collective Intelligence

Segment 1: Introduction to Participatory Media and Collective Intelligence

00:00 [MUSIC PLAYING]

00:09 SIMON LINDGREN: H
professor of sociology
am an internet resear
media and how peopl
communicate and org
both political and cultu
about participatory me

00:31 SIMON LINDGREN [c
During the last few de

The Prosumer and the Project Studio: The Battle for Distinction in the Field of Music Recording
Steven James Cole
In: *Sociology* (2011)
ARTICLE — SAGE journals

Identifying the Events That Connect Social Media Users: Charting Follower Accession on Twitter
In: *SAGE Research Methods Cases* (2014)
CASE — SAGE researchmethods

10 Differentiating Practices of Participatory Visual Media Production
Richard Chalfen
In: *The SAGE Handbook of Visual Research Methods* (2011)
BOOK CHAPTER — SAGE researchmethods

Social Media Explosion
Marcia Clemmitt
In: *CQ Researcher* (2013)
REPORT — CQ RESEARCHER

Social Media and Pol
Tom Price

anyway because the authors are using *enough* of the same language, even if they diverge sometimes.

The real limitation of this approach becomes apparent when there is not enough textual content to compensate for the occasional false negative or false positive in this way. SAGE's content is an excellent case study, because as well as books, articles, cases, reports, and other full-text content, we have a growing amount of multimedia and numerical content, as well as a quantity of PDF-only journal backfiles. In fact, in terms of citable objects (to which one might assign a DOI for instance), only around 20% of SAGE's online content can be considered traditional text-based documents.

This challenge, however, is not an insurmountable one because "nontext" content such as multimedia and numerical data sets also

Although students are accustomed to watching videos for their classwork and course work, many are unaware of resources that are available through their library: In a previous SAGE white paper on students and video in higher education, we found that less than a third of students we surveyed had ever searched for videos via their institution's library (Leonard, 2015). With most students using YouTube as their first, and sometimes only, port of call, we believe that there is a tremendous untapped benefit in linking academic and educational videos to other types of scholarly and educational content.

On the page shown in Figure 2, a user can view a tutorial-style video on participatory media and collective intelligence. The video's keywords have been mapped to SAGE's social science thesaurus, enabling easier comparison across all SAGE's content types. For example, one concept introduced in the video is "prosumption," and if the viewer is interested in advanced research on that concept, then the SAGE Recommends tab links to an article discussing presumption in the academic journal literature. Alternatively, the viewer may be interested in more introductory context on how participatory media is produced and find the chapters from SAGE Research Methods of interest. More general still is a report from CQ Researcher contextualising the growth of social media as a whole. Fittingly, the naturally small semantic fingerprint of short videos adds an element of serendipitous chaos and the viewer can continue to scroll through various other texts and numerical data, with the widget picking up on all aspects of the content that the viewer is currently engaged with to provide a variety of recommendations rather than simply a list of virtually identical encyclopaedia entries that the viewer might have already seen in their first set of search results.

often has text-based metadata available. What text there is tends to be scarce (for example abstracts and subject metadata), so it is crucial to make that small amount of text content as effective as possible. In SAGE Recommends, we employ a *knowledge-based* approach on top of the string-based approach. In addition to creating tf-idf vectors from strings of text, we create tf-idf vectors for concepts extracted against a controlled vocabulary: In short, we teach the natural language processing software about the strings in the text and how to treat them. Extracting concepts against a thesaurus allows much more control of the terms that are extracted, in three ways:

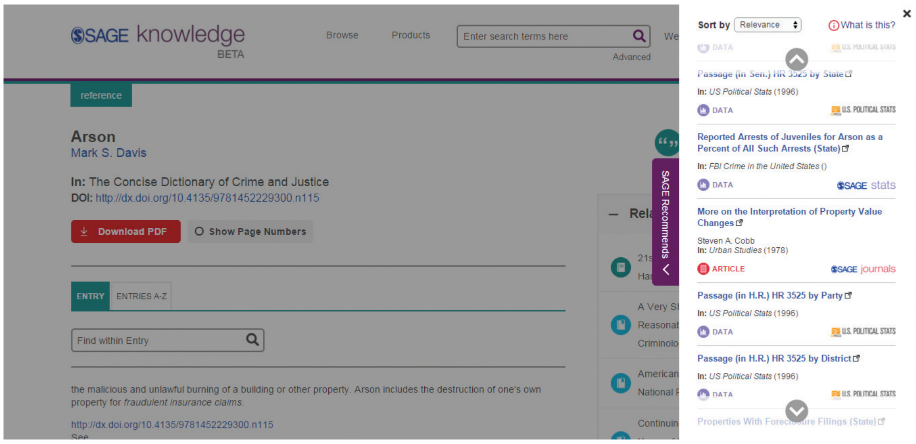
- It allows for the creation of rules for disambiguating identical strings. For example, competitive video game teams and ancestral groups can be extracted as different concepts despite using the same string (“clans”).
- It allows for the creation of rules for combining different strings. For example, we can state that “marijuana” and “cannabis” are the same concept.
- It allows for the creation of an authority list of terms that are considered important and meaningful to a given discipline. For example, *acceleration* is a precise term in engineering and physics but could be considered to be a generic term in the social sciences, where it might lead to spurious similarity. This is particularly useful where a high degree of control is needed because little text content is present.

A typical numerical data series or video may only have a title or half a dozen keywords by way of textual content, but by making sure that each of those keywords maps to the same vocabulary that is used to annotate all content in the same discipline, and by giving a statistical boost to documents that share these important concepts, that small amount of textual content can go a long way (see Figure 3).

This solution can also be considered a new problem for organizations that do not already have ontologies or vocabularies that describe their content. SAGE was no different, particularly as most of our content is in the social sciences, where there is no equivalent to canonical vocabularies, such as the National Library of Medicine’s Medical Subject Headings (MeSH), that can be adopted. The good news is that building huge information resources such as ontologies and thesauri can be assisted by text-mining techniques similar to those used for processing text into numerical fingerprints. SAGE, for instance, drew on our 50 years of academic publishing as the starting point for sketching a map of the social sciences, mining the implicit and explicit structure built into our encyclopaedias’ headwords and reader’s guides as a shortcut to a hierarchy of important concepts on a given topic, and by extension the whole of the social sciences. Supplemented by public domain vocabularies, and hundreds of hours of human intervention in refining the term list and developing the rule base, the resulting 63,000-concept vocabulary is not only a statement of the domains in which SAGE publishes but also a valuable information resource that can be used to more accurately tag and index content and improve all modes of discovery, including the serendipitous.

Content is arguably *the* core competency of a publisher, and it may become crucial that any publisher is able to develop comprehensive resources that describe and represent its content. By drawing out the implicit structure in their content and making it explicit, information providers can derive intelligence from their own content to drive better relevance, context, and insight. Here semantic enrichment is a virtuous circle, with text mining facilitating the creation of vocabularies and thesauri that can, in turn, be used to further enrich content more accurately.

Figure 3 SAGE Recommends linking to numerical data



Arson
Mark S. Davis

In: The Concise Dictionary of Crime and Justice
DOI: http://dx.doi.org/10.4135/9781452229300.n115

the malicious and unlawful burning of a building or other property. Arson includes the destruction of one's own property for *fraudulent insurance claims*.

http://dx.doi.org/10.4135/9781452229300.n115

Passage (in Sen.) HR 3525 by State
In: *US Political Stats* (1996)
DATA — U.S. POLITICAL STATS

Reported Arrests of Juveniles for Arson as a Percent of All Such Arrests (State)
In: *FBI Crime in the United States* ()
DATA — SAGE stats

More on the Interpretation of Property Value Changes
Steven A. Cobb
In: *Urban Studies* (1978)
ARTICLE — SAGE journals

Passage (in H.R.) HR 3525 by Party
In: *US Political Stats* (1996)
DATA — U.S. POLITICAL STATS

Passage (in H.R.) HR 3525 by District
In: *US Political Stats* (1996)
DATA — U.S. POLITICAL STATS

Properties With Foreclosure Filings (State)

This is one possible way to design for serendipity, among many, but it appears to be working to deliver serendipitous recommendations to our users at the point of need. By making the item that is currently being viewed the point of discovery, the recommendation is to some extent depersonalized, and the emphasis is on the content itself, allowing the user to focus on the next point on their current discovery journey. This is not to say that we have skewed the aforementioned balance between interestingness and relevance; as we observed, recommendations can be less useful if they are obvious, and so SAGE Recommends uses a number of strategies to ensure the novelty and interestingness of discoveries while still not deviating the user too far from their current path.

An important design decision was to ensure that no single product dominates in the recommendations. This is a danger for all publishers with a large volume of a particular type of content, such as SAGE's backfiles of around 1.5 million journal articles. Including all of that content at once creates a risk that users will be recommended journal articles and nothing else, which seems to be antithetical to the principle of serendipity. Some publishers have approached this problem by providing a federated-style search results page, providing different sets of results for different types of content, but this again can encourage users to stay in their comfort zones.

If readers were to alight on the "Arson" entry in The Concise Dictionary of Crime and Justice on SAGE Knowledge (see Figure 3), they could use the page not just as a means to quickly understand what is meant by the concept but also as a jumping-off point for further discovery. If users were to run another search for "arson" on SAGE Knowledge, they would be presented with content very similar to what they are currently viewing: chiefly reference and books content giving an overview of the subject. SAGE Recommends goes beyond this and shows the reader content that they might not even have known existed, in particular statistics from the FBI Crime in the United States series on SAGE Stats, showing arrest rates for arson and giving readers the ability to compare the prevalence of arson across the United States. The widget also recommends data from a House floor vote on redefining arson, which again is information that users might not necessarily have been aware of. The spirit is to tempt users to take their research into interesting, new directions while maintaining high levels of relevance so that they are not taken too far away from their current information need.

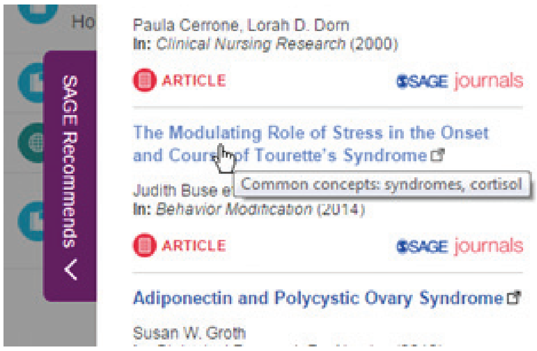
We chose to solve this problem by identifying documents from these large corpora that were considered to be more problematic for semantic enrichment (e.g., metadata-only content or abstract-only content), in effect cherry-picking just those journal articles most susceptible to this approach. This has the effect of not only improving the relevance of the journal articles that are recommended, but by standardizing the size of the corpus across different products, a wide mix of different products and content is also more likely.

By artificially restricting the size of some corpora, the effect is to move the dial away from relevance and towards novelty and interestingness. It also encourages the user to leave their comfort zone and explore content that they might not otherwise have searched for, or indeed even be aware existed. One aspect of serendipity is to discover something unintentionally, but this also speaks to an aspect of serendipity that is often forgotten: that serendipity should be unexpected. Users would expect that SAGE publishes books and journals but may be surprised to find data and multimedia available to them.

But serendipity is more than just a chance encounter: There also needs to be a discovery, an insight. If the first aspect of serendipity can be engineered behind the scenes, this second step has to be engineered (if indeed it can be engineered at all) in the user interface. We did a lot of testing with users to establish exactly what kind of information they need to quickly assess and ascertain the relevance of a piece of information. We used this as part of the rationale for designing the widget as something that overlays the page, as recommendation systems that try to fit all relevant information into a small section of the body of the page inevitably omit essential metadata (and sometimes even the complete title of a recommendation). In expanding a collapsible widget, the user gives tacit permission to take up as much of the screen as is needed to display the metadata necessary to make sense of a recommendation, which our testing established is, at a minimum, title, author, date, source, and type of content.

Often this is not enough, so we also go beyond this to expose the rationale of the recommendation to users. If users hover their mouse over a link, they are shown the concepts that the link has in common with what they are currently reading. Rather than simply telling users, “This is related to that,” we say, “This is related to that because it also talks about this.” This approach allows for a speedy evaluation of the link, especially when a nondescriptive title is used for the document, such as “Introduction” or “Editor’s Foreword”). It also allows for the quicker formation of a potential insight. In the screenshot shown in Figure 4, users might wonder why SAGE Recommends has shown them a journal article on Tourette syndrome when they are reading about Cushing syndrome. If they hover over the link, they will see that “cortisol” is linked to both disorders, which is a link that users might not have otherwise had the time to make.

Figure 4 Common concepts between two documents



Paula Cerrone, Lorah D. Dorn
In: *Clinical Nursing Research* (2000)

The Modulating Role of Stress in the Onset and Course of Tourette's Syndrome

Common concepts: syndromes, cortisol

Judith Buse et
In: *Behavior Modification* (2014)

Adiponectin and Polycystic Ovary Syndrome

Susan W. Groth

We have discussed that this solution is based around SAGE’s users and SAGE’s content. Although the extensibility of this approach to serendipity in the larger scholarly environment is hopefully apparent, the current limitation of this approach is that it does not link out to content from other publishers or to open resources.

There is an obvious tension between the need for publishers to develop workflows, solutions, and resources that map closely to their own content and the need to interoperate with other organizations and their resources. Until the shape of the landscape becomes clearer, it may be that the best individual publishers can do is to be able to move quickly to interoperability (which is why our thesaurus has dormant crosswalks to other vocabularies), and in the meantime make it clear to users that single-publisher solutions are not intended to promise an authoritative method of discovery but rather a list of suggestions for further exploration.

References

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*(13), 13-18.

Leonard, E. (2015). *Great expectations: Students and video in higher education* (White paper, p. 12). Thousand Oaks, CA: SAGE Publishing. Retrieved from https://us.sagepub.com/sites/default/files/studentsandvideo_0.pdf

Maloney, A., & Conrad, L. Y. (2016). *Expecting the unexpected: Serendipity, discovery, and the scholarly research process* (White paper). Thousand Oaks, CA: SAGE Publishing. Retrieved from https://us.sagepub.com/sites/default/files/SerrDiscovery.pdf