

On Deriving Indicators from Texts

Steven O. Kimbrough¹, Thomas Y. Lee¹, and Ulku Oktem¹

University of Pennsylvania, Philadelphia PA 19104, USA

Abstract. This paper presents and explores the idea of deriving numerical indicators from texts, that is, converting text data to numerical data that has predictive or diagnostic value. One application of such a general capability is to the provisional identification of networks, or rather, of associations within networks. Conversely, given a network structure among entities that are associated with various texts, the network structure can itself contribute usefully to construction of indicators derived from texts. The focus of the paper is on basic concepts and methods for deriving indicators from texts. Much research remains to be done.

Key words: text mining, text data mining, data mining, mashing, economic indicators, social indicators

1 A First Look

It is fun to experiment by counting term frequencies using search engines. One way of doing this is to go to Google's Advanced News Search at <http://news.google.com> and undertake queries to compare monthly changes in word usage for the news sources indexed by Google. For example, Google reports "about 644 [hits] from Jun 1, 2006 to Jun 30, 2006 for [the search term] recession" and "about 4,440 from Jul 1, 2006 to today [July 30, 2006] for recession." So talk of recession was up, it would seem. We would like to have more than two months of data, of course, but Google presently only supports these queries for the current and previous months. With diligence over time or recourse to other sources, this is a problem that can be overcome.

A more interesting problem is whether counting the appearances of search terms signifies anything important at all. To take the present case, looking back from April 2008, did the uptick in mentions of "recession" and related terms provide evidence that a recession was coming soon? Nouriel Roubini, professor of economics at New York University and economics commentator for Bloomberg News, thought it did (<http://www.rgemonitor.com/blog/roubini/137589>).

Nouriel Roubini — Jul 24, 2006

It is hard to predict with certainty whether the U.S. and global economy will suffer of serious stagflation or even a recession (my bearish views are fleshed out in my recent blogs here and here). I have been arguing that those risks are large and rising; and I have recently argued that the probability of a US recession in 2007 is, in my view, as high as 50%. In brief, the Three Bears of high oil prices, rising inflation leading to higher policy rates, and a slumping housing markets will derail the Goldilocks (of high growth and low inflation) and trigger a sharp U.S. slowdown in 2006, that may turn into a recession in 2007.

One potential barometer of such recession concerns - with all the appropriate caveats - is how many news articles are citing terms such as stagflation, U.S. recession, or recession in general. Here is a brief news-mood summary taken from a brief search on Google News today:

- News Citations of Stagflation: 655
- News Citations of US Recession: 2,870

– News Citations of Recession: 4,850

And it is not just obscure publications that are worrying about stagflation and recession. Recent detailed discussions of such risks were recently front page on the WSJ and on Bloomberg. And the number of private sector folks, experts and academics talking about such risks is rising. The authoritative Mike Mussa, former Chief Economist at the IMF, now puts the odds of a US recession at 25-30% while the Fed's own internal yield curve model now predicts that the probability of a U.S. recession in 2007 is almost 40%. As the proverb says, talk is cheap (if so sweet) but in this case the evidence that many folks and leading media publications are increasingly and systematically talking about recession and stagflation to the tune of 1000s of recent articles and commentaries should be at least a signal, to policy makers and market folks, that these risks may be rising (and the talk is no [sic; so? not?] sweet).

These numbers may not mean much, taken by themselves, since we do not have any base rate information. Perhaps, after all, discussion of stagflation declined over the previous year.

Let us look then at some other queries on Google. On July 30, 2006 the number of hits is

- “about 134 from Jun 1, 2006 to Jun 30, 2006 for stagflation” and “about 438 from Jul 1, 2006 to today for stagflation”
- “about 7,270 from Jun 1, 2006 to Jun 30, 2006 for inflation” “about 39,200 from Jul 1, 2006 to today for inflation”
- “about 6,130 from Jun 1, 2006 to Jun 30, 2006 for disaster” and “about 29,800 from Jul 1, 2006 to today for disaster”
- “about 2,620 from Jun 1, 2006 to Jun 30, 2006 for global-warming” and “about 10,500 from Jul 1, 2006 to today for global-warming”
- “about 611 from Jun 1, 2006 to Jun 30, 2006 for biofuel — bio-fuel — biofuels — bio-fuels” and “about 2,890 from Jul 1, 2006 to today for biofuel — bio-fuel — biofuels — bio-fuels”
- “about 7,250 from Jun 1, 2006 to Jun 30, 2006 for peace” and “about 75,400 from Jul 1, 2006 to today for peace”
- “about 1,260 from Jun 1, 2006 to Jun 30, 2006 for prosperity” and “about 7,080 from Jul 1, 2006 to today for prosperity”
- “about 7,640 from Jun 1, 2006 to Jun 30, 2006 for baseball” and “about 92,600 from Jul 1, 2006 to today for baseball”

So, it would seem, not only were disasters and recessions upon us, things were getting hotter, too, and there was even much greater interest in peace, prosperity and baseball. One strongly suspects internal bias here. Given that we do not know how Google keeps the records and given that *every* search term increases strongly, one strongly suspects that Google was aging its document collection. Hits were up recently in all categories because the great majority of documents on hand are comparatively more recent.

Still, if Google is not at least straightforwardly a reliable source, perhaps there are others. The blogger Kevin Drum had a more nuanced take on Professor Roubini's findings (http://www.washingtonmonthly.com/archives/individual/2006_07/009250.php).

Drum's table looks better, although there will always be skeptics. This was posted two minutes later in comment on Drum's entry.

Yes, but how many of those are articles pondering the number of times the word “stagflation” is used in other articles?

Posted by: Uncle Vinny on July 28, 2006 at 1:45 PM — PERMALINK

POLITICAL ANIMAL

By Kevin Drum

July 28, 2006

STAGFLATION....The *Washington Post* reports that the latest economic news is grim:

The nation's gross domestic product, which measures the value of all goods and services produced, rose at a below-average 2.5 percent annual rate in the second quarter....Meanwhile, consumer prices shot up at a heated 4.1 percent annual pace.

The combination of slow growth and high inflation, of course, is "stagflation," and yesterday [Brad DeLong](#) linked to a [Nouriel Roubini](#) piece suggesting that the number of news reports mentioning stagflation (a "potential barometer" of recession) had been quite high recently.

But is that true? Is the number not just high, but *higher than usual*? Only a chart can tell us for sure! And here it is: the number of citations of the word "stagflation" from Nexis over the past year. (The July 2006 number is a projection.)

Sure enough, Roubini is right: mentions of stagflation spiked heavily starting last month. And given today's news, I'll bet they'll spike even higher in the coming months. If news cites really are a decent way of projecting economic performance, the news is not good.

—Kevin Drum 1:43 PM [Permalink](#) | [Trackbacks](#) | [Comments \(90\)](#)

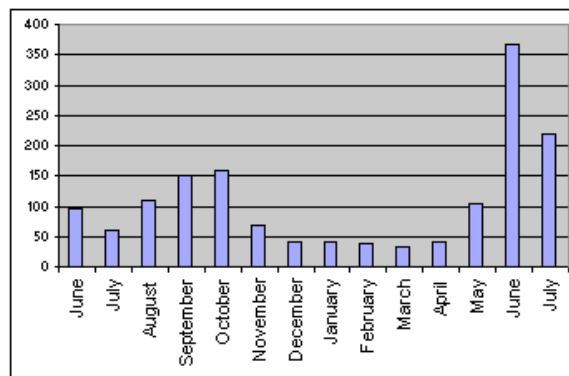


Fig. 1. Kevin Drum on "stagflation"

If, however, we take a somewhat longer view, using Factiva (more or less equivalent to LexisNexis) and querying on "stagflation", we get Table 1. While this is broadly in agreement with Drum's findings on LexisNexis, the longer history of the Factiva data provides no credible support for the hypothesis that in July 2006 there had been a significant uptick in discussion of stagflation. Instead, what we have looks like merely a typical two- or three-month flare-up, of which there was a stronger one in April and May of 2005. Even this is far from established as a pattern. The fact is, these term-count data, these indicators derived from texts, provided no reason whatsoever to believe that stagflation or recession was just around the corner. Which is just as well, since it wasn't. Note, however, that in late 2007 and early 2008 the U.S. economy did experience significant slowdown, if not an actual recession. The numbers for this period are in fact much elevated compared to the earlier data.

The term-count data just described are examples of *indicators derived from texts*. In the examples above, the indicators are defined simply by the number of search results over a collection of text documents

Month	Count	Month	Count
January 2005	43	October 2006	73
February 2005	44	November 2006	99
March 2005	51	December 2006	57
April 2005	364	January 2007	47
May 2005	244	February 2007	37
June 2005	72	March 2007	62
July 2005	57	April 2007	187
August 2005	87	May 2007	86
September 2005	145	June 2007	37
October 2005	178	July 2007	52
November 2005	63	August 2007	37
December 2005	53	September 2007	108
January 2006	41	October 2007	77
February 2006	44	November 2007	260
March 2006	55	December 2007	534
April 2006	38	January 2008	530
May 2006	76	February 2008	945
June 2006	247	March 2008	666
July 2006	151	April 2008	327
August 2006	227	May 2008	740
September 2006	140	June 2008	

Table 1. “stagflation” document counts on Factiva

lection of documents, as on display above, is simple, even naïve. It would be quite surprising if this alone could produce a great deal of information worth attending to. Thus, our second issue is what we will call the *aptness question* for indicators: What are the more apt—more useful—indicators that may be constructed from texts, and how might Web services facilitate their construction? These two issues are not unrelated, but let us discuss them separately.

2 The Validity Question

Indicators derived from texts are a form of quantitative data. As with any data, it is appropriate to ask what, if anything, they are able to predict, either alone or in combination with other data. And, as with any other data, the validation of predictive powers is something for which there are standard, well-established methods, broadly statistical in nature and well known to the empirical sciences. In this regard, data derived from texts has no special status, and indeed is no different than data obtained by any other method. So, while it is inappropriate to notice an uptick in news reports mentioning stagflation and conclude that stagflation is on

(Google, Nexis, Factiva). In the past, constructing such an indicator would be anything but simple, requiring manual searching and counting or programming using a customized, source-specific application programming interface (API). Today, the construction of simple indicators based upon aggregating search results over topics and time is facilitated through the use of Web services. Google,¹ LexisNexis,² and Factiva³ all export standardized Web services to facilitate searching and aggregation for such tasks as the construction of indicators.

As fully admitted, it is an open question whether these indicators actually provide information on anything besides themselves. So our first issue might be parsed as Given an indicator, what does it indicate, with what strength, and how is this known? We’ll call this *the validity question* for indicators. Assuming it is possible to get past the first issue, the validity question, the second issue concerns the creation of indicators from texts. Constructing indicators using common term counts from a general and unrefined col-

¹ <http://code.google.com/apis/ajaxsearch/>

² <http://www.lexisnexis.com/webserviceskit/developers/>

³ <http://factiva.com/competitiveintelligence-services.html>

the way, it is quite appropriate to *test* whether indicator data obtained from news reports can actually be used to make useful predictions.

There is a complication and it applies to all quantitative data. To illustrate, reconsider the stagflation example. You have decisions to make for which the state of the economy in the next few years will have great impact. It matters a great deal to you whether or not stagflation conditions will obtain, and you seek to discern the future. Moreover, the situation is far from clear. The well-tested indicators are not conclusive, yet you must make decisions. Suppose it is true that there has been a recent uptick in news stories about baseball. Are you interested for the sake of your present decisions? Probably not. What about the uptick in stories mentioning stagflation? Probably yes, but not because the uptick leads you to believe that stagflation is on the way. Rather, the uptick creates a *prima facie* reason to attend to it and possibly to make an effort to obtain further information. In the case of an indicator derived from texts, you might consider reading the texts to see what they say.

The larger point is that there is more than one reason why an indicator can be valid, can be worth paying attention to. First, an indicator may be a statistically valid predictor of a condition of import. If so, we may say that the indicator is *statistically valid*. The upshot of our original discussion of word counts of “stagflation” on Google News was that it is a non-starter to consider this indicator to be statistically valid (on the evidence to hand). Second, an indicator may be valuable because it is thought (with justification) to be credibly associated with information that substantially bears on the decision to hand. Such an indicator may be called *investigationally valid*. In brief, while word counts of “stagflation” might not be statistically valid, they well could be (and probably are) investigationally valid in some circumstances. Having an interest in the future state of the economy and noticing a surge of news reports, opinion pieces, blogs, etc. mentioning stagflation, it will often be entirely reasonable to pay attention and to investigate further. And this is where indicator data derived from texts may be different: you can read the documents that produced the data, and you may learn something that will, with warrant, affect your decision.

Indicators and indexes that aggregate indicators (and are themselves indicators) are widely published and used, although it is rare for them to be derived from texts. Familiar, or prominent, examples include:

- Economic indicators. The federal government publishes vast numbers of economic indicators.⁴ There are private collectors of indicators.⁵ The Conference Board publishes a widely-cited Index of Leading Economic Indicators as well as its Global Business Cycle Indicators.⁶
- Commerce-related indicators
 - Ranking products and services based upon attributes drawn from user-generated online reviews. (See for example [AGI07, LHC05, LLW08, SBC⁺07].)
 - Financial, or investment, analysis using text-based indicators seeks to derive indicators for such commercially interesting events as bankruptcies, risk, and changes in profitability. See [BGI08, Cec05, Lee08, Li06].
- Environmental indicators. There are sustainability indicators.⁷ There are Environmental Treaties and Resources indicators.⁸ And others.

⁴ <http://www.economicindicators.gov/> and <http://www.gpoaccess.gov/indicators/index.html>

⁵ <http://www.economic-in-dicators.com/>, or <http://www.rogerseconomics.com/Indicators/index.html>

⁶ <http://www.conference-board.org/> and <http://www.conference-board.org/economics/bci/>

⁷ <http://www.sustainablemeasures.com/>

⁸ <http://sedac.ciesin.columbia.edu/entri/>

- Social indicators. The World Bank publishes social indicators of development.⁹ The United Nations publishes various social indicators.¹⁰ as well as its Millennium Development Goals Indicators.¹¹ Thompson publishing offers its own Essential Science Indicators(TM).¹²

With regard indicators, we may distinguish:

1. What it is a given indicator (or statistic) is (purportedly) about, the event or condition about which the indicator bears information. So, for example, unemployment levels may indicate something about present or future economic growth; similarly the number of building permits issued in the past month may indicate something about levels of housing construction in the near future.
2. How well an indicator indicates. An indicator for X may or may not carry much information about X.
3. How interesting or useful an indicator is. A very weak indicator may be interesting because what it is about is important and no stronger indicators are available.

Generally, indicators are published because they are thought to be useful and important, not because they have been proved to be statistically valid. Validation is typically left to the user. Indexes, however, are often assembled using statistical modeling techniques and are claimed by their creators to have a degree of statistical validity (e.g., the various economic indexes, which aim to predict performance of particular economies). Typically, these indicators are used both in modeling exercises, in which statistical validity is sought, and less formally as inputs to decision processes, which judge them to be investigatively valid.

As seen in Section 1, one can construct indicators and indexes by composing one or more Web services. In the same way, an indicator or index may itself be exported as a Web service. In particular, we can integrate an indicator directly into a statistical tool for checking validity, as input for constructing another index, or into a decision support tool.

In the remainder of this paper, we focus on investigational validity for indicators derived from texts. This should be seen as a precursor to studies of statistical validity. We suspect that there is ample contribution to decision making to be made on the investigational front, and we wish to explore it a bit here. Where applicable, we will point out how indicators can be constructed from and exported as Web services. For now, we will consider the second issue before us.

3 The Aptness Question

Investigationally valid indicators command our attention for decision making. Mere occurrence counts of “stagflation” and “recession” on Google or Factiva may or may not. The aptness question may be stated as follows: Under what conditions and in what sorts of ways can indicators—indicators derived from texts—be obtained that warrant attending to during decision making? This is a large question, certainly too large to be treated definitively in any single paper, let alone in a paper that introduces the question. Our present aim is more modest. By presenting examples of plausibly apt indicators derived from texts, we demonstrate that the aptness question has favorable answers. There are ways, as we shall show, of deriving indicators from texts, which indicators merit consideration during decision making. Here, then, are a few examples.

⁹ <http://www.ciesin.org/IC/wbank/sid-home.html>

¹⁰ <http://unstats.un.org/unsd/demographic/products/socind/>

¹¹ <http://mdgs.un.org/unsd/mdg/default.aspx>

¹² <http://scientific.thomson.com/products/esi/>

Year	Count
1990	6262
1991	4493
1992	8074
1993	3724
1994	3541
1995	4057
1996	4687
1997	16827
1998	13777
1999	11422
2000	17582
2001	17514
2002	13271
2003	10753
2004	21817
2005	33880
2006	58591
2007	129958
2008*	45770

Table 2. All sources ("global warming") on Factiva. (*Document counts for 2008 are for 1 January – 19 June)

Forecasting economic events, with real money on the line, is as difficult a prediction task as there is. Any simple way of predicting economic trends or other events that would affect stock prices can be expected to be arbitrated away quickly. Predicting other kinds of events and conditions will often be easier. The aptness of deriving indicators from texts may well be favorable for more promising tasks. For example, searching Factiva on "global warming" yields the results shown in Table 2.

It is evident that the number of recorded mentions of "global warming" has increased substantially since 1990 and that an important transition was made between 1996 and 1997. We do not know how much of the increase in the counts of articles mentioning global warming is attributable to Factiva's collection policies. Perhaps the periodicals indexed have changed greatly since 1990. Other possibilities are conceivable but implausible: Perhaps the average periodical has gotten longer. *Prima facie*, taking the data at face value, however, simple document occurrence counts on "global warming" seem to have increased about 20-fold between 1990 and 2008, and there seems to have been a significant uptick during the 1996-7 period. If you are concerned with the amount of "play" that global warming is getting in the popular press—perhaps because you engaged in strategic planning or public relations for a petroleum company—then it is not unlikely that you would find the data in Table 2 investigatively interesting.

A legitimate worry about the data in Table 2, just noted, is the possibility that the underlying document base varies so as to produce misleading results. One way to address this worry is to select a fixed number of periodicals that existed and were indexed throughout the

time period in question. Table 3 Left reports the number of articles containing "global warming" and appearing in either *The Wall Street Journal* or *The Washington Post*.

Comparing the results in Tables 2 and 3 Left we see that they are broadly in agreement: during 1990–2 a baseline is established, during 1993–6 a new, lower baseline obtains, a strong uptick occurs in 1997, followed by a general increase until the present. In short, there is remarkable agreement.

The term "sustainability" has become focal for many interested in environmental issues, broadly construed. In ordinary language and in its recent environmental context, "sustainable" as applied to a system or process connotes maintainability over a long period of time because a renewable balance has been achieved. The fact of global warming, together with the fact that it is caused in large part by such "greenhouse" gases as carbon dioxide and methane, strongly suggests that unconstrained burning of fossil fuels (petroleum and especially coal) is not sustainable. Heating of the earth will halt it, working either by human foresight and planning or by shutting down civilization, if not our species. The principal policy goals of what may be called the sustainability movement are described in this representative passage (<http://www.anavogroup.com/sustainability.html>, accessed 31 July 2006):

The 'sustainability' movement strives to bring technologies and processes to the marketplace that support both basic human needs, and often socially accepted comforts, while being environmentally benign and economically viable.

The same site asserts that “Sustainability is Gaining Traction”:

Today, sustainability perspectives and innovations are becoming mainstream at a dizzying pace.

From the data, it would indeed appear that global warming is “gaining traction.” What does a similar investigation say of sustainability? Tables 3 and 4 supply supporting evidence. Table 3 reports on the counts of articles in *The Wall Street Journal* or *The Washington Post* that mention “sustainability.” The pattern broadly resembles that for “global warming”: there is an appreciable uptick in 1997 and continued growth from there. Table 4, left, repeats the query, now with all the news sources that Factiva indexes, with quite similar results. Finally, Table 4, right, queries all news sources on Factiva for “sustainable development,” a phrase closely associated with the sustainability movement. The usual pattern roughly reasserts itself.

It is apt to count occurrences of “global warming,” “sustainability,” and so on if your aim is to discover whether associated topics are “gaining traction.” For some purposes and for some decisions, the evidence presented here would constitute sufficient warrant to judge that both global warming and sustainability are increasingly topics of general concern. For other purposes and decisions, the evidence would not suffice, but would be interesting nonetheless because it presents a *prima facie* case, whose significance would trigger further investigation. Such investigation might take many forms, including deriving other indicators from texts, and actually reading indicated documents, as well as more traditional forms such as undertaking public surveys.

Year	Count
1990	294
1991	189
1992	238
1993	102
1994	81
1995	78
1996	82
1997	268
1998	224
1999	169
2000	242
2001	337
2002	221
2003	169
2004	318
2005	467
2006	803
2007	1386
2008*	583

Year	Count
1990	16
1991	16
1992	14
1993	12
1994	15
1995	12
1996	11
1997	34
1998	18
1999	29
2000	32
2001	41
2002	69
2003	95
2004	101
2005	88
2006	117
2007	155
2008*	83

Table 3. Left: *The Wall Street Journal* or *The Washington Post* ((rst=J or rst=WP) and "global warming") on Factiva. (*Document counts for 2008 are for 1 January – 19 June.) Right: *The Wall Street Journal* or *The Washington Post* ((rst=J or rst=WP) and "sustainability") on Factiva. (*Document counts for 2008 are for 1 January – 19 June.)

Year	Count	Year	Count
1990	376	1990	649
1991	527	1991	781
1992	891	1992	2316
1993	1195	1993	2069
1994	1833	1994	3580
1995	2188	1995	3298
1996	3291	1996	4292
1997	5688	1997	6155
1998	7226	1998	6839
1999	9405	1999	8797
2000	11329	2000	9693
2001	13013	2001	11195
2002	22118	2002	26659
2003	27154	2003	17626
2004	34102	2004	20180
2005	40259	2005	23852
2006	51298	2006	25380
2007	76410	2007	33758
2008*	43296	2008*	16052

Table 4. Left: All sources ("sustainability") on Factiva. Right: All sources ("sustainable development") on Factiva. (*Document counts for 2008 are for 1 January – 19 June.)

4 Further Methods

The previous section, §3 on aptness, constitutes a basic demonstration of the aptness, at least in some circumstances, of deriving indicators from texts. The methods we employed—essentially calculating hit counts of news articles (counts of documents returned by the searches) and categorizing the hits by year—were quite elementary. Our purpose in this section is to indicate some additional methods that may be used for effectively—aptively, to coin a neologism—deriving indicators from texts.

The aptness of deriving indicators from texts may, as we have just argued, be enhanced by judicious choice of topic. It may also be enhanced by careful selection of source documents. Factiva is just one of very many potential sources of documents and its focus, broadly news articles, is narrow. Other sources of interest include:

1. Other commercial document indexing and retrieval services, such as LexisNexis and ABI-Inform.
2. Organizational archives and repositories, including technical reports and working papers.
3. General Web queries with search engines such as Google and Yahoo!
4. Specialized Web queries, e.g., using Google Scholar.
5. Directed Web queries, e.g., using `wget` given a list of URLs.
6. Regulatory files, e.g., SEC filings (10-K, 10-Q, etc.), EPA and OSHA filings, and their analogs outside the United States.
7. Patents and patent applications (both US and non-US).

Systematic exploration of all methods and sources is well beyond the scope of this, or any single paper. We content ourselves here with examples to illustrate our basic points. However, even a few examples are sufficient to illustrate how Web services may also support additional approaches for deriving indicators from text. Next, we discuss some useful methods, using a document base of 800,000+ US patents, published between 1999 and 2004.

4.1 Categorization of Indicators

A very useful way of generalizing, or abstracting, the aptness examples in §3 is to see the reports, presented in the various tables, as examples of arrays of triples, each consisting of the *value* of an *indicator* in a *category*. For example, in the first row of Table 3 (page 8) we find: 1990 and 16. The indicator is the count of articles containing “sustainability”. The category is a complex one (double in this case): appearing in 1990 and in either *The Wall Street Journal* or *The Washington Post*. And the value is 16. Notice that there is a many-to-many relationship between indicators and categories. In Table 3 one indicator’s values are shown for several categories (the year changes). Conversely, we might fix the year and the periodicals (constituting the category) and vary the indicators, using, for example, “biofuels”, “water resistant” and so on.

Table 5 illustrates the generalization. Here the indicator is the count of documents containing the term biodegradable. The categories are US patents published in the period 1999-2004 × USPTO classification class. We see that classes 424 and 514 of the USPTO classification scheme have apparently identical class descriptions (they in fact do, weirdly, as do 606 and 604).

Notice that the table displays a *pattern* of information: values of the indicator are displayed across multiple categories. This *pattern-oriented* query—showing a non-random distribution of values for multiple categories—is in distinction to the *record-oriented* queries familiar to users of Internet search engines. Record-oriented queries return lists of documents; pattern-oriented queries return scores (data values) for categories of documents. See [DKP00] for discussion of the distinction between record-oriented and pattern-oriented queries.

Patterns of indicators can be constructed by composing multiple Web services together. In this case, we can combine data from news sources and the USPTO classification scheme into a denormalized relational structure using a service like Yahoo’s YQL (<http://developer.yahoo.com/yql/>) Multi-dimensional queries can then produce some classes of pattern-oriented results [DKP00].

4.2 Calculating More Nuanced Indicators

So far, the indicators we have discussed have been functionally trivial: they are simple summations of term counts. While there is much to be said in favor of simplicity, there is no reason not to use complex functions and multiple parameter inputs if that proves useful. To illustrate, the following is a small step in that direction.

Figure 2 is an example outcome which indicates the number of “energy efficiency” related patents that have been granted to the indicated companies between 1999 and 2003 and their relevancy to the topic. The relevancy score is obtained as follows:

Relevancy = ((Total number of times “energy efficiency” is mentioned in patents where there is at least one mention of “energy efficiency”)/(Total number of patents where “energy efficiency” is mentioned at least once)) × 10 (scaling factor).

The Figure 2 graph indicates that although Air Products has the most patents related to “energy efficiency” among the top chemical companies interested in this concept, Kimberly Clark’s patents are more strongly connected to “energy efficiency”. Similarly, the most popular energy efficiency topics covered in these patents are found very quickly but not included in this document.

Class Num.	Count	Class Description
424	1638	Drug, bio-affecting and body treating compositions
514	1624	Drug, bio-affecting and body treating compositions
435	1063	Chemistry: molecular biology and microbiology
510	563	Cleaning compositions for solid surfaces, auxiliari
623	415	Prosthesis (i.e., artificial body members), parts
606	358	Surgery
604	280	Surgery
428	275	Stock material or miscellaneous articles
536	233	Organic compounds – part of the class 532-570 ser
530	221	Chemistry: natural resins or derivatives; peptides
210	211	Liquid purification or separation
528	187	Synthetic resins or natural rubbers – part of the
430	159	Radiation imagery chemistry: process, composition,
600	137	Surgery
525	117	Synthetic resins or natural rubbers – part of the
524	111	Synthetic resins or natural rubbers – part of the
264	106	Plastic and nonmetallic article shaping or treatin
47	91	Plant husbandry
523	79	Synthetic resins or natural rubbers – part of the
134	79	Cleaning and liquid contact with solids
427	78	Coating processes
508	74	Solid anti-friction devices, materials therefor,
8	74	Bleaching and dyeing; fluid treatment and chemical
252	69	Compositions
800	68	Multicellular living organisms and unmodified part
162	68	Paper making and fiber liberation
106	63	Compositions: coating or plastic

Table 5. Counts of US patents, 1999-2004, mentioning “biodegradable” by USPTO classification scheme classes, in descending order

In this particular example, nuanced indicators come not only from document searches but also through text processing of each document in the search result. A number of Web services supporting text mining (see <http://u?compare.org/> and <http://www.alchemyapi.com/>) enable the analysis of individual text documents and larger text collections. The process for constructing more nuanced indicators composes a text processing Web service with a document search Web service.

4.3 Ontologies

An ontology, in the currently popular sense of term as deployed in the information sciences, is “a specification of a conceptualization” (accessed 2006-08-20: <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html>). To elaborate:

A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among themA conceptualization is an abstract, simplified view of the world that we wish to

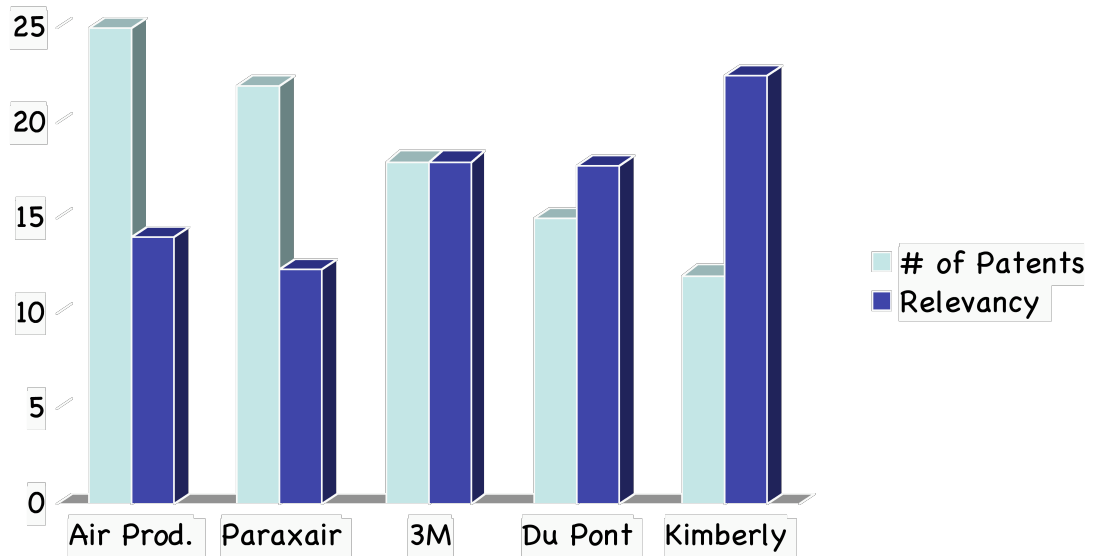


Fig. 2. Two measures of “energy efficiency” in patents

represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly.

An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what “exists” is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory.

The author adds in a footnote that

Ontologies are often equated with taxonomic hierarchies of classes, but [sic; with?] class definitions, and the subsumption relation, but ontologies need not be limited to these forms. Ontologies are also not limited to conservative definitions, that is, definitions in the traditional logic sense that only introduce terminology and do not add any knowledge about the world

Thus, a taxonomic classification system may constitute a form, perhaps a minimal form, of ontology. The work we describe in this paper pertaining to ontologies is in the main focused on use of taxonomic classification systems. There is, however, no fundamental reason to so limit the employment of ontologies for deriving indicators from texts. It is merely a matter of convenience for present purposes. Drawing on expertise from a number of sources (including one of the authors of this paper), we have constructed a rudimentary ontology

(*qua* taxonomic classification system) on the subject of biofuels in the context of sustainability. Our biofuels ontology currently encompasses about 600 n-grams (words or word phrases) as leaves in a taxonomic hierarchy. The ontology continues to evolve; in consequence specific reports discussed here may not fully reflect its current, more complete, state.

ngram	count(ngram)
methanol	52653
biodegradable	10222
ethanol	3355
viscosity	3015
recyclable	2909
energy efficient	2240
solar energy	1978
air pollution	1947
fuel efficiency	1822
emulsions	1182
methane	894
increase productivity	739
water pollution	483
water usage	393
energy reduction	214
emulsification	207
greenhouse gases	197
wind energy	194
transesterification	158
fuel pump	124
higher energy efficiency	102
waste reduction	95
viscosity reduction	33
water efficiency	18
reduction in usage	9
environmentally sustainable	5
lower water usage	4
organic certification	4

Table 6. N-grams from the biofuels ontology and their frequencies in US patents, 1999-2004

Table 6 presents representative n-grams from our ontology along with their frequency counts in US patents from the 1999-2004 period. Notice that ‘methanol’ appears much more than any other biofuel n-gram, strongly suggesting that many of its occurrences are unrelated to biofuels. Table 7 probes this notion by looking at co-occurrences of biofuels n-grams. Here we see that 5,367 patents contain ‘vegetable oil’ and of these 1,996 contain ‘methanol’, suggesting a fairly strong association between vegetable oil and methanol in the context of biofuels. The nature of that association must, of course, be discovered by other means, such as reading in the patent documents themselves.

One of the services an ontology provides is to draw our attention to topics of import, and how they are commonly described, in the subject to hand (here, biofuels). This is one way in which the expertise embodied in the ontology is leveraged. We note that this motivation is also present in the concept of *exemplary*

ngram	doc. count
methanol	1996
biodegradable	520
ethanol	101
emulsions	65
viscosity	63
recyclable	31
water pollution	26
air pollution	21
methane	17
emulsification	10
solar energy	10
transesterification	9
energy efficient	8
used vegetable oil	6
increase productivity	6
fuel efficiency	5
water usage	3
wind energy	3
waste vegetable oil	2
greenhouse gases	1

Table 7. Counts in n-grams in US patents, 1999-2004, among the 5,367 documents containing ‘vegetable oil’

documents developed by Blair and Kimbrough [BK02]. Their notion might be summarized as a concept for building ontologies by identifying key or important—their word is *exemplary*—documents in an area and then to using information retrieval methods—particularly *seed searching*—to exploit the ontologies. This is a promising idea which remains to be investigated rigorously.

A number of general purpose ontologies are publicly exported via Web services for use in constructing indicators and indices. Among them, a Wikipedia API (unofficially at <http://www.programmableweb.com/api/wikipedia> and an official API coming soon via http://wikimedialab.org/en/index.php/Wikipedia_API) and the forthcoming DMOZ API for Open Directory (<http://blog.dmoz.org/bloggers/bob-keating>).

4.4 Information Mashing

Information mashing is a new term of art, referring to the aggregation of information from multiple sources to serve a common purpose. The earliest very clear articulation that we have found of the term and the idea is by Ellen Miller of the Sunlight Foundation (www.sunlightfoundation.com) in her blog on April 28, 2006 (<http://www.sunlightfoundation.com/node/465>). She writes:

Information Mashing. Don’t you just love that term? It’s one of the major goals of Sunlight and while we’ve been working on it for the past couple of months we have a ways to go before it happens in any substantial way. Our goal is simple: integrate in a user-friendly way individual data sets (like campaign contributions, lobbyists and government contracts) that makes the whole larger than the sum of its parts.

We’d like to create something we’ve dubbed an “Accountability Matrix.” A website where, with one click you can look up a major donor and see not just their campaign contributions, but also

their lobbying expenditures, the names of members who've flown on their private jet, the names of former congressional staffers they've hired, and so on.

In a nutshell, we want to make information more liquid and more accessible to the public.

She has politics and public policy specifically in mind. Generalizing this is the notion of a vaim, a value-added information mash. (See <http://opim.wharton.upenn.edu/~sok/asadai/vaim-faqs.pdf> and <http://opim.wharton.upenn.edu/~sok/asadai/vaim-design-faqs.pdf>.) The vaim concept generalizes to any subject area of interest (e.g., sustainability) and explicitly adds the notion of adding value through leading edge software technology. The key point, however, is the generalization of what she calls the “Accountability Matrix.” This is essentially a way of associating information items in a useful way. Her example is a good one. There are documents (including Federal Election Reports) about campaign contributions and there are documents (e.g., press releases, newspaper stories) about hiring of former congressional staffers. What Miller is asking for is a system that would facilitate the explicit association of (certain types of) information items.

The general principle is that fact *A* and *B*, that *A* is associated with *B*, may well be much more significant than either the fact *A* or the fact *B* in isolation. We call this the *pattern principle* or the *connecting the dots principle*. For further discussion, see [DKP00]. A vaim is a system that works to implement the pattern principle, using documents from multiple sources to make explicit associations among information items.

Here, to illustrate, is a simple example of information mashing. We have built a second ontology, this one for firms active in the biofuels space. By extracting assignee information from the patents and using string-matching queries, we can use the company ontology to obtain patenting information regarding the biofuels companies. Which companies hold patents? In what years? What are the n-grams present? In addition, we downloaded all of the publicly-available Web pages at each company's site and indexed them with the original biofuels n-grams. In consequence we have two document collections—patents and company Web pages—each indexed with the same two ontologies—biofuels n-grams and biofuels companies. This effects a ‘mashup’ and affords comparison of patent topics and word patterns with Web page topics and word patterns. Table 8 presents a small example.

As befitting a Web-based phenomenon like mashups, there are a number of Web services supporting the construction of mashups. Among them, Yahoo Pipes (<http://pipes.yahoo.com/pipes/>) and Google's Mashup Editor, which has been superseded by the App Engine (<http://code.google.com/appengine/>), are well known. Commercial tools such as those supported by IBM's Mashup Center 2.0 are also available (<http://www?01.ibm.com/software/info/mashup-center/>).

Organization	N-gram	count(ngram)
Biodiesel Industries	biodegradable	9
Biodiesel Industries	greenhouse gases	3
Biodiesel Industries	air pollution	2
Biodiesel Industries	water pollution	1
Biodiesel Industries	recyclable	1
BioDiesel International	biodegradable	5
BioDiesel International	recyclable	1
BioDiesel International	greenhouse gases	1
Biodiesel Technologies Inc.	greenhouse gases	2
Biodiesel Technologies Inc.	air pollution	1
Biodiesel Technologies Inc.	biodegradable	1
Biofuels Corporation plc	biodegradable	5
Biofuels Corporation plc	greenhouse gases	3
Biofuels Corporation plc	air pollution	1
Cargill	biodegradable	15
Cargill	waste reduction	4
Cargill	increase productivity	3
Cargill	organic certification	3
Cargill	environmentally sustainable	1
Cargill	greenhouse gases	1
Cargill	water usage	1
Detroit Diesel Corporation	fuel efficiency	57
Detroit Diesel Corporation	recyclable	1
Filter Speciality Inc.	air pollution	2
Filter Speciality Inc.	biodegradable	1
Filter Speciality Inc.	recyclable	1
Greenline Industries	biodegradable	1
Greenline Industries	recyclable	1
Neste Oil	biodegradable	3
Neste Oil	recyclable	1
Neste Oil	air pollution	1
Patriot Biofuels	biodegradable	2
Patriot Biofuels	fuel efficiency	1

Table 8. Example n-gram patterns by firm

4.5 Information Extraction

Information extraction is a term of art in the information sciences:

Information extraction (IE) is a type of information retrieval whose goal is to automatically extract structured or semistructured information from unstructured machine-readable documents. It is a sub-discipline of language engineering, a branch of computer science.

(http://en.wikipedia.org/wiki/Information_extraction, accessed 2006-8-20; see [JM02] for a summary of the state-of-the-art.) What has gone by the name of information extraction does not include the derivation of indicators from texts that we describe here. Instead, IE has focused on extracting particular facts from texts; it has been, as we say, record-oriented rather than pattern-oriented. See, e.g., [CKL04] for an example of using information extraction techniques to find uses for products from patent documents. [GT94] is an example of deriving hazardous events data from news sources, but the extraction was done manually and would not be classified today as *information extraction*. This said, deriving indicators from texts is surely a form of information extraction in a broad, ordinary language sense.

Further, it is certainly the case that information extraction and deriving indicators from texts are complementary. We used basic forms of information extraction on the patent documents to obtain such data as dates, assignees, and classifications, which we used to obtain many of the results described above. Our view is that any and all methods that prove useful are welcome for converting texts into data and deriving indicators. We merely observe that these methods are not presently well collected. Some are standard in the IE community, some in the broader information retrieval community, some are new (including some of our techniques) and originate outside either community. Together, we believe, these techniques can be very effectively deployed for deriving indicators from texts.

4.6 Association Distributions

The derived indicators we have discussed so far have all been scalar quantities, mainly document counts, or some richer function, organized by category. Vector indicators consisting of multiple scalar indicators have been investigated and found effective for information discovery (in Dworman's Ph.D. thesis, [Dwo99a]; see [DKP00] for an overview). In particular, Dworman investigated categorized patterns of term co-occurrences. We can illustrate this in our present context, with patent documents and ontologies for biofuels and biofuel firms.

The biofuels firm Cargill was found to have 136 patents in our patent document base. Table 9 shows the distribution of biofuels n-grams for Cargill patents. Detroit Diesel, another biofuels firm, has 102 patents, but a very different n-gram profile, as Table 10 shows. Finally, we can compare these firms with patenting from MIT. During the 1999-2004 period MIT scored 657 patents, resulting in the biofuels n-gram distribution shown in Table 11.

ngram	count(ngram)
viscosity	63
methanol	44
biodegradable	41
vegetable oil	36
transesterification	27
ethanol	21
landfill	16
emulsions?	13
waste water	9
methane	8
waste disposal	6
water treatment	5
collectors?	3
energy efficient	2
waste water treatment	2
sustainable	2
emulsification	2
emulsions	1
waste management	1
water supply	1
fresh water	1
greenhouse gas	1
pesticides?	1

Table 9. Biofuels n-gram patterns for Cargill. 136 patents in all. Note: “pesticides?” matches to “pesticide” and “pesticide” and similarly for other “s?” constructions. See also Tables 10 and 11.

ngram	count(ngram)
fuel pumps?	9
fuel efficiency	9
viscosity	3
methanol	2
heating systems?	2
emulsions?	1

Table 10. Biofuels n-gram patterns for Detroit Diesel. 102 patents in all.

ngram	count(ngram)
ethanol	89
methanol	75
viscosity	63
emulsions?	58
biodegradable	47
collectors?	15
methane	12
energy storage	10
sustainable	7
fuel efficiency	6
solar energy	6
emulsification	5
heating systems?	3
fuel pumps?	2
pesticides?	2
drinking water	2
water supply	2
energy efficient	2
space heating	1
waste disposal	1
energy ratios?	1
pollution prevention	1
waste minimization	1
waste water	1
landfill	1
municipal waste	1
waste production	1
water treatment	1
energy balance	1
transesterification	1
vegetable oil	1

Table 11. Biofuels n-gram patterns for MIT. 657 patents in all.

Finally, two points about association patterns. First, the vectors introduce an additional aspect of pattern-oriented retrieval. Now we compare and assess different distributions (vectors) consisting of multiple scalar values. Second, Tables 9–11 *illustrate* the method. They hardly exhaust its variations. Our indicator-category-value framework remains apt, but now our indicators are vector (or distribution) patterns. This forecloses no option regarding creation of more complex categories. For example, we might compare n-gram distributions among (assignees \times patents containing the n-gram ‘vegetable oil’). This is the level of complexity investigated by Dworman, although in entirely different subject domains. We have illustrated above use of a simpler category; obviously much more complex ones are available should they prove useful.

5 Categorized Document Bases with Normalized Hits

In a CDB (categorized document base [KMRT07]) the documents in the collection are tagged with taxa from classification systems. The documents tagged with a common taxon may be said to form a subcollection. If we query a CDB we can return counts of documents (numbers of ‘hits’) by category or subcollection that match our query. These hit numbers may be of considerable interest.

Assume, for example, that we have a document collection pertaining to biofuels in which each of the documents is mapped to one or more firms from the S&P 500. Querying the collection results in matches between the query and a number of documents in the collection. It is then possible to count the number of matching documents (i.e., the ‘hits’) by subcollection or taxon, that is by firm from the S&P 500.

The utility of having such hit information for a query is substantial in principle. If our collection of documents is interesting and relevant to a given topic and the query is of interest for a given purpose, then the hit counts may have investigational (or stronger) validity for comparing distinct taxa. The hit counts are not, however, comparable with each other unless they are normalized in an appropriate way across the various subcollections, each constituted by the documents associated with a given taxon.

This normalization can be done in any of a number of ways. Perhaps the most obvious is to obtain a ‘raw’ hit count score and modify it by a factor representing the relative size (in any of a number of senses, including number of documents, number of words, and so on) of the subcollection to hand. A second approach, which would appear to be equally valid, is to construct the subcollections in such a way that they are sized more or less equally. Then with equal-sized normalization we can compare the raw hit counts directly. It is impossible to know a priori which normalization will be most useful in a given situation. Only future experience will be dispositive on this matter.

To illustrate, we used Yahoo! to identify and download a collection of about 1000 biofuels-related documents (973 were actually obtained). We then mapped each firm in the S&P 500 to the 30 most relevant documents from the collection (for that firm). Interestingly, this mapping resulted in the use of 946 (of the 973) documents. No document was associated with more than 40 firms and only 3 documents were associated with more than 30 firms.

The result was *not* equal-sized subcollections. Only 326 of the S&P 500 firms hit at all. Of these, 203 firms/taxa had fully 30 associated documents in their subcollections. Another 24 had between 20 and 29 associated documents. Reducing the cutoff point further, 248 firms had more than 10 associated documents.

Even though not every firm/taxa was represented in the subcollections and the subcollections were of differing sizes, the matching can be considered a soft normalization. All but 80 of the S&P 500 firms were either well represented (more than 10 associated documents) or were absent entirely from the matching. Thus, approximately and we think good enough for many investigatory purposes, the represented firms were represented equally. Corresponding subcollections are roughly comparable for purposes of noticing unusual numbers of hits.

Querying the collection on ‘glycerol’ (an important by-product of making biodiesel) and ordering the hit categories in descending order of hit intensity produces the information in Table 12. Querying on ‘transesterification’ (an important process in the making of biodiesel) produces the information displayed in Table 13.

Results are easily proliferated with the existing system (biofuels documents stored in a database, indexed for full text retrieval and indexed by S&P 500). Beyond ‘glycerol’ and ‘transesterification’ any reasonable query can be issued, and the resulting data obtained, along with the underlying documents. Beyond this, the approach applies to any collection of documents to which a classification scheme is matched in the manner described. We submit that with such a system, the bar for investigational validity of these indicator data is readily surpassed.

Figures 3–7 illustrate the concept of CDBs with Normalized Hits in a deployed prototype system, called Sizatola Categorization Engine. In these figures we see demonstrated the mashing (matching) of a classification system and a document collection, and some of the resulting possibilities thereby created. Here the classification is the UNSPSC (Standard Products and Services Classification, www.unspsc.org) system, which consists of some 22,000 taxa and aims to classify all of the world’s products and services. We obtained 10 documents for each of the UNSPSC taxa, thereby creating a CDB suitable for equal-sized normalization.

UNSPSC is a four-level classification system. The figures illustrate a query sequence that navigates the hierarchy. In Figure 3 the user queries the top level of the hierarchy (consisting of separate *segments*, in the argot of UNSPSC) for high viscosity. As we see, only one segment is implicated, Chemicals including Bio Chemicals and Gas Materials. There are 629 hits. Drilling down to the family level (below segment), we see in Figure 4 that only one family is implicated: additives. Drilling down further, Figure 5 shows a rather extensive list of additives with varying degrees of hit levels. Clicking on mud removal mixtures takes us to the commodity level at which there are two commodities present; see Figure 6. Finally, we see in Figure 7 the result of drilling down to the leaf level by clicking on mud cleanout agents. There are just two documents (hits) available. The user may view the hits as stored or may go to the original URL and explore further.

With just a few quick clicks the user is able to exploit the hierarchical classification system, as mapped to the document collection, and the pattern of hits created by the query in order to focus attention on very specific, and often quite surprising, topics.

Score	Company
4	Jack Henry & Associates
4	Ultra Petroleum
4	United States Steel
3	Ameron International
3	PAETEC Holding
3	Manhattan Associates
3	SI International
3	SRA International
3	Cleveland-Cliffs
3	Hewitt Associates
3	Brinker International
3	Burlington Northern Santa Fe
3	American Tower
3	International Speedway
3	Wendy's International
3	Crown Castle Intl
3	Lennox International
3	CAI International
3	Bucyrus International
3	NII Holding
2	Endo Pharmaceuticals
2	Dollar Thrifty Automotive Group
2	Arch Coal
2	Spirit Aerosystems Holdings
2	Coeur d'Alene Mines
2	Automatic Data Processing
2	Adobe Systems
2	Computer Sciences
2	Valeant Pharmaceuticals
2	Armor Holdings
2	Trinity Industries
2	Leisure time
2	Watson Pharmaceuticals
2	Armstrong World Industries
2	Hovnanian Enterprises
2	Cavco Industries
2	Mohawk Industries
2	Reliance Steel & Aluminum

Table 12. Top match scores on 'glycerol' for 'biofuels' documents, by S&P 500 company

Score	Company
8	Alexander & Baldwin
6	Pioneer Natural Resources
6	United States Steel
5	PAETEC Holding
5	Ultra Petroleum
5	Palm
5	American Tower
5	NII Holding
4	Penn Virginia
4	Burlington Northern Santa Fe
4	King Pharmaceuticals
4	Manhattan Associates
4	Jack Henry & Associates
4	Watson Pharmaceuticals
4	Hewitt Associates
4	Cleveland-Cliffs
4	Sigma Designs
4	Chevron
3	Global Payments
3	ConocoPhillips
3	Panera Bread
3	Barr Pharmaceuticals
3	Endo Pharmaceuticals
3	Arch Coal
3	Stage Stores
3	Valeant Pharmaceuticals
3	Hovnanian Enterprises
3	Crown Castle Intl
3	Adams Respiratory Therapeutics
3	Sykes Enterprises
3	Werner Enterprises
3	Reliance Steel & Aluminum
3	Forest Laboratories
3	Automatic Data Processing
3	Silicon Laboratories
3	Foundation Coal Holdings
3	Steel Dynamics
3	Hewlett-Packard

Table 13. Top match scores on 'transesterification' for 'biofuels' documents, by S&P 500 company

The screenshot shows a web browser window titled "Sizatola Categorization Engine". The address bar contains the URL "http://opim.wharton.upenn.edu/~edkim2/dcapture/home2.php". The browser's search bar is empty, and the page title is "Sizatola Categorization Engine".

The main heading is "Sizatola Categorization Engine". Below it, the "Search Terms:" section contains two input fields with the text "viscosity" and "high". There is a button labeled "add another search term".

The "Start Date:" and "End Date:" fields are empty. The "Limit:" field is set to "10".

There are two buttons: "Run Query!" and "Clear Page".

The search results are displayed in a table with the following data:

Segment	Count	Family	Link
Chemicals_including_Bio_Chemicals_and_Gas_Materials	629	family	link

Fig. 3. SCE: Simple AND Query (*viscosity* and *high*) with UNSPC Taxonomy and 10 Documents per Taxon, Using the Drill Down Interface at the Top Level (Segment) of the UNSPC Taxonomy

The screenshot shows a web browser window titled "Sizatola Categorization Engine" with the URL <http://opim.wharton.upenn.edu/~edkim2/dcapture/home2.php#>. The search terms "viscosity" and "high" are entered in the search box. Below the search box are buttons for "add another search term", "Run Query!", and "Clear Page". There are also dropdown menus for "Start Date", "End Date", and "Limit".

Family	Count	Class
[Chemicals_including_Bio_Chemicals_and_Gas_Materials]		
Additives	629	class

Segment	Count	Family	Link
Chemicals_including_Bio_Chemicals_and_Gas_Materials	629	family	link

Fig. 4. SCE: Simple AND Query (viscosity and high) with UNSPC Taxonomy and 10 Documents per Taxon, Using the Drill Down Interface at the Second Level (Family) of the UNSPC Taxonomy, under the Segment `Chemicals_including_Bio_Chemicals_and_Gas_Materials`

The screenshot shows a web browser window with the URL <http://opim.wharton.upenn.edu/~edkim2/dcapture>. The page title is "Sizatola Categorization Engine". The search terms entered are "viscosity" and "high". The interface includes fields for "Start Date", "End Date", and "Limit", along with "Run Query!" and "Clear Page" buttons.

Class	Count	Commodity
<i>[Additives]</i>		
Colloids	98	link
Surfactants	75	link
Fluid_loss_additives	73	link
Paraffin_asphaltene_control_agents	46	link
Curing_agents	45	link
Friction_reducers	43	link
Polymer_breakers	38	link
Emulsion_breakers	34	link
Clay_stabilizers	34	link
Anti_sludgers	34	link
Plasticizers	15	link
Corrosion_inhibitors	15	link
Gas_hydrate_controllers	11	link
In_situ	11	link
Oil_well_sealants	10	link
Buffers	8	link
Mud_removal_mixtures	8	link
Anti_gas_migration_agents	8	link
Anti_oxidants	5	link
Expanding_agents	5	link
Scale_controllers	4	link
Bactericides	3	link
Extenders	2	link
Retarders	2	link
Flame_retardants	1	link
Chemical_scavengers	1	link
Family		
<i>[Chemicals_including_Bio_Chemicals_and_Gas_Materials]</i>		
Additives	629	class
Segment		
Chemicals_including_Bio_Chemicals_and_Gas_Materials	629	family link

Fig. 5. SCE: Simple AND Query (*viscosity* and *high*) with UNSPC Taxonomy and 10 Documents per Taxon, Using the Drill Down Interface at the Third Level (Class) of the UNSPC Taxonomy, under the Segment *Chemicals_including_Bio_Chemicals_and_Gas_Materials* and under the Family *Additives*

The screenshot shows a web browser window with the URL <http://opim.wharton.upenn.edu/~edkim2/dcaptur>. The page title is "Sizatola Categorization Engine".

Search Terms:
 viscosity
 high
 add another search term

Start Date: [] End Date: []
 Limit: []

Run Query! Clear Page

Commodity	Count	Link
[Mud_removal_mixtures]		
Mud_removal_mixtures	6	link
Mud_cleanout_agents	2	link

Class	Count	Commodity
[Additives]		
Colloids	98	link
Surfactants	75	link
Fluid_loss_additives	73	link
Paraffin_asphaltene_control_agents	46	link
Curing_agents	45	link
Friction_reducers	43	link
Polymer_breakers	38	link
Emulsion_breakers	34	link
Clay_stabilizers	34	link
Anti_sludgers	34	link
Plasticizers	15	link
Corrosion_inhibitors	15	link
Gas_hydrate_controllers	11	link
In_situ	11	link
Oil_well_sealants	10	link
Buffers	8	link
Mud_removal_mixtures	8	link
Anti_gas_migration_agents	8	link
Anti_oxidants	5	link
Expanding_agents	5	link
Scale_controllers	4	link
Bactericides	3	link
Extenders	2	link
Retarders	2	link
Flame_retardants	1	link
Chemical_scavengers	1	link

Family	Count	Class
[Chemicals_including_Bio_Chemicals_and_Gas_Materials]		

Fig. 6. SCE: Simple AND Query (viscosity and high) with UNSPC Taxonomy and 10 Documents per Taxon, Using the Drill Down Interface at the Fourth Level (Commodity) of the UNSPC Taxonomy, under the Segment Chemicals_including_Bio_Chemicals_and_Gas_Materials and under the Family Additives and under the Class Mud_removal_mixtures

The screenshot shows a web browser window titled "Sizatola" with the URL <http://opim.wharton.upenn.edu/~edkim2/dcapture/hor>. The browser's address bar and tabs are visible, including "Gmail: Email from Google" and "UNSPSC Homepage".

Sizatola Categorization Engine

Search Terms:

Start Date: End Date:

Limit:

Mud_cleanout_agents

[\[cached\] \[URL\] 11_www.window.state.tx.us\\$taxinfo\\$audit\\$oilwell\\$service\\$glossary.htm](#)
[\[cached\] \[URL\] 13_www.worldoil.com\\$TechTables\\$Fluids_Desc.asp](#)

Commodity	Count	Link
<i>[Mud_removal_mixtures]</i>		
Mud_removal_mixtures	6	link
Mud_cleanout_agents	2	link
Class	Count	Commodity
<i>[Additives]</i>		
Colloids	98	link
Surfactants	75	link
Fluid_loss_additives	73	link
Paraffin_asphaltene_control_agents	46	link
Curing_agents	45	link
Friction_reducers	43	link
Polymer_breakers	38	link
Emulsion_breakers	34	link
Clay_stabilizers	34	link
Anti_sludgers	34	link
Plasticizers	15	link
Corrosion_inhibitors	15	link
Gas_hydrate_controllers	11	link
In_situ	11	link
Oil_well_sealants	10	link
Buffers	8	link
Mud_removal_mixtures	8	link
Anti_gas_migration_agents	8	link
Anti_oxidants	5	link

Fig. 7. SCE: Simple AND Query (*viscosity* and *high*) with UNSPC Taxonomy and 10 Documents per Taxon, Using the Drill Down Interface at the Fourth Level (Commodity) of the UNSPC Taxonomy, under the Segment *Chemicals_including_Bio_Chemicals_and_Gas_Materials* and under the Family *Additives* and under the Class *Mud_removal_mixtures* and under the Commodity *Mud_cleanout_agents*

6 A Governing Framework and Related Work

Categorized document bases are hardly new, even if the term CDB is. What is particularly new is our emphasis on their role in text mining. We view text mining, or knowledge discovery in text (KDT), as an effort to discover new knowledge. In distinction, information retrieval is about finding relevant documents. Heretofore, categorization of text has mainly been used to support retrieval, rather than—as emphasized here—discovery.

Retrieval succeeds if the documents identified in response to a query are in fact (comparatively) highly relevant to the query, or to the intended goal of the user who has formulated the query. (For present purposes we ignore this important distinction.) Discovery succeeds if we find patterns of information that are surprising and warrant follow on investigation. In short, patterns that are investigatively valid.

A general framework will help articulate this idea and serve to point to related work. We call this the C-M-A framework for knowledge. How in general may knowledge—interesting, investigatively valid patterns of information—be obtained from collections of text (as well as other information items)? In the C-M-A framework there are three salient steps

1. Categorization/classification
2. Measurement
3. Association

which we discuss individually beginning in the next section.

In the context of the C-M-A framework (or process), our philosophical view is that interesting, investigatively valid patterns of information *constitute* knowledge. To discover interesting, investigatively valid patterns of associations among categorized measurements simply *is* to discover knowledge. We allow that there may be many ways of discovering knowledge, but this is surely among them.

6.1 Categorization/classification

A CDB is created by categorizing (classifying, applying a taxonomy to, etc.) the documents in a collection. A number of ways are available, discussed below, for categorizing documents. Also, more than one classification scheme may be applied to a body of documents. Indeed, faceted indexing systems, such as the Getty's Art & Architecture Thesaurus, consist of multiple classification systems, each called a facet (http://www.getty.edu/research/conducting_research/vocabularies/aat/; see also

http://www.getty.edu/research/conducting_research/vocabularies/).

Documents are routinely classified both by manual methods and by automated methods. Among the automated methods, *text categorization*, *document clustering* and *automated indexing* name the principal families of approach.

Document classification by supervised learning methods, referred to as *text categorization*, has been extensively researched in the context of information retrieval and, to a lesser extent, text mining. Good overviews are available in [FS07, JM02, Kon06].

Document classification by unsupervised methods, also known as *document clustering*, has likewise been extensively researched in the context of information retrieval and, to a lesser extent, text mining. Good overviews are available in [FS07, JM02, Kon06].

Automatic indexing (also *automatic classification* and *automatic document classification*) names a body of method and research arising from the information and library science communities [Moe00, SWY75]

that has aimed in large part at algorithmic creation of document indexes. This literature has focused almost exclusively on information retrieval. Exploring its relevance for KDT presents an exciting prospect.

We note that *sentiment analysis* and *information extraction* [FS07, JM02, Kon06] name two important and very active areas of research and development in information retrieval and in text mining. *Sentiment analysis* is “A technique to detect favorable and unfavorable opinions toward specific subjects (such as organizations and their products) within large numbers of documents [it] offers enormous opportunities for various applications. It would provide powerful functionality for competitive analysis, marketing analysis, and detection of unfavorable rumors for risk management.” (Accessed 24 August 2006: http://www.tr1.ibm.com/projects/textmining/takmi/sentiment_analysis_e.htm.) See [HL04, Lee04, LHC05, NY03, YNBN03, PE05]. They are, or have been, less central to document classification. In the future, however, it is entirely possible that these techniques could make important contributions here as well.

Manual methods for document classification predated by far all of the automated methods and have been studied in depth in the information science literature. In a typical regime, a *controlled vocabulary* (aka: classification system, taxonomy, etc.) is created and indexers are trained to read documents and apply the vocabulary terms as appropriate. This has been found to be an expensive and not terribly reliable process.

Interestingly, manual classification has come back in force with the fielding of *tagging* systems seen at such sites as del.icio.us (<http://del.icio.us/>), Flickr (<http://www.flickr.com/>), and Google’s email system (<http://mail.google.com/>). These and other applications have spawned the *folksonomy* concept, a folksonomy being a taxonomy created from the bottom up, by the folk.

Folksonomy (also known as collaborative tagging, social classification, social indexing, and social tagging) is the practice and method of collaboratively creating and managing tags to annotate and categorize content. In contrast to traditional subject indexing, metadata is generated not only by experts but also by creators and consumers of the content. Usually, freely chosen keywords are used instead of a controlled vocabulary [Vos07]. Folksonomy is a portmanteau of the words folk and taxonomy, hence a folksonomy is a user generated taxonomy. (<http://en.wikipedia.org/wiki/Folksonomy>, accessed 2008-6-22)

Invention of the term is credited to Thomas Vander Wal, who maintains a related blog and Web site (<http://www.vanderwal.net/folksonomy.html>, accessed 2008-6-22). A popular article in *Salon* in 2005 stimulated much interest in folksonomies [Mie05]. “Folksonomies – Cooperative Classification and Communication Through Shared Metadata” by Adam Mathes has been highly regarded in this community (<http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>, accessed 2008-6-22).

And there is the inevitable entry from *Wired*, “Order out of chaos” by Bruce Sterling (<http://www.wired.com/wired/archive/13.04/view.html?pg=4>, published April 2005, accessed 2008-6-22).

The considerable hoopla about folksonomies has generated inevitable questioning of their usefulness (*D-Lib Magazine* is a good source, recently <http://www.dlib.org/dlib/november06/peterson/11peterson.html>). Even so, a major folksonomy software project is underway in the museums and archives world, called Steve (<http://www.steve.museum/>). For those interested in KDT, developments in folksonomies bear monitoring.

All of these approaches are potentially useful for deriving indicators from text. Whether, or how well, they work can only be determined by experience and careful empirical investigation. The illustrations presented above use mainly other methods. For example, patent documents are published in categorized form. Today they typically use XML and a raft of special tags for mark up and categorization. To create a CDB

one does what we did: process the documents to capture the categorization information (date, patent number, inventors, etc.), store the categorization information in a relational database, and index the body of the documents with a convenient full text system, preserving the associations between the document's text and its categories.

The illustrations in §5 used two different methods. First, we obtained a collection of documents on a particular subject (here: biofuels) and then we mapped classification systems to the collection by taking the n most relevant documents for each taxon in the classification scheme. This method draws upon and is described in [KMR07]. Second, we used the taxa of a classification system to retrieve relevant documents (in this case from the Web). The resulting collection was categorized by the taxa simply because each document in the collection was causally obtained by some taxon or other. This method draws upon and is described in [KMRT07].

6.2 Measurement

In information retrieval we apply an algorithm to a query and a *collection* of documents in order to produce a relevance score for each *document*. The score may be binary 1-0 (relevant or not) for matching retrieval (by document) or graded by relevance on a finer scale, in the case of relevance ranking retrieval (by document).

In KDT we apply an algorithm to a query and a *CDB* in order to produce a relevance score for each *category* (or combination of categories). Again, the score may be binary 1-0 (relevant or not) for matching retrieval *by category* or, much more likely, graded by relevance on a finer scale, in the case of relevance ranking retrieval *by category*.

In short, what fundamentally distinguishes information retrieval from text mining are the objects of retrieval. In the case of information retrieval the objects are individual documents, for which we seek relevance scores. In the case of text mining the objects are categories of documents (categorized sub-collections of documents), for which we seek relevance scores.

We are not aware of much research on measuring *category* relevance to a query. In the examples we presented above we largely relied on counts of matches to a query produced by a pattern-oriented retrieval engine (typically a boolean search engine). In some cases we reported the counts directly, in others we reported functional transformations.

It is easy to imagine other forms of measurement for category relevance. One might, for instance, use the average or the maximum of the relevance rank scores for the documents in a category. Or one might treat all of the documents in a category as one large document and use standard IR methods on the resulting document set. Various forms of query expansion might be applied, either to the query or to the process of identifying the documents to be in a given category. This is an important area for future investigation.

6.3 Association

By *quantity* let us mean the value of a particular measurement on a category. Thus, for example, if the category is plasticizers and 15 is the measurement value on this category for the query *high AND viscosity* (see Figure 7), then we say that 15 is the quantity (of the query *high AND viscosity* on plasticizers). Note that quantities are inherently numeric. This is without loss of generality. Even nominal data (e.g., country of residence, state, gender, marital status) may be coded numerically. Specific quantities may be defined on nominal, ordinal, interval or ratio scales. The quantities in our examples in this paper are all ratio scaled.

For the sake of discovering knowledge we compare different quantities in order to determine whether or not there is an association between (or among) them. Does one co-vary in a non-random fashion with another? If so, then the two quantities have a non-random association or, to be brief, simply an association.

And if there is an association, then one quantity indicates the other (to some degree or other). (The point generalizes to groups of quantities.)

In sum, given quantities—categories and numerical measurements for them—produced by applying a scoring engine to a query on the CDB, we have demonstrated the deriving data (or indicators) from text. We have reduced the text mining problem to the more familiar realm of data exploration and analysis. In the examples above we used simple tabular and graphical means to display data (derived from text) in one and two dimensional forms. More complex forms are described in [KMRT07] and merit further investigation. Related ideas are presented in [LLP07].

The fundamental point here is that the concepts and methods we have described for converting documents to data are fully general and open up to text mining the richly endowed world of statistics and data mining. As a general fact, knowledge is discovered by discovering non-random associations among quantities. Our point is that categorization and measurement operations on collections of text provide suitable and productive quantities, comparison of which will often lead to knowledge discovery.

7 Discussion

Broadly speaking, we would like to make two points. First, we have observed that data may be—indeed very often are—valuable indicators for decision making even in the absence of established statistical validity. If data are known to be statistically reliable indicators, that is certainly welcome and to be sought after. The press of events, however, will often preclude decision making in the presence of this kind of statistical certainty. This is well accepted. Investigational aptness is often available even if statistical validity is not. What is done is to employ subjective judgment (e.g., “A rising divorce rate would seem to indicate that families are under increasing stress and traditional norms are weakening”) and a preponderance of indicators (divorce rate, suicide rate, labor participation rate, ...) to weave together a plausible story to support action. Social science, public policy, and works by “public intellectuals” largely fit this form. Robert Putnam’s *Bowling Alone* [Put00], Francis Fukuyama’s *Trust* [Fuk95], and Eric Beinhocker’s *The Origin of Wealth* [Bei06] are just three of many excellent examples.

Our second point, to which most of the paper is devoted, is to demonstrate that useful data—data that can serve as indicators for decision making—can be derived from bodies of texts in a variety of ways. The examples we have presented include well-known and established techniques, new techniques, and combinations thereof. Moreover, these techniques are supported by any number of existing or emerging Web services that both standardize and simplify the construction of indicators and indexes. We have assumed on the basis of face validity that there is a *prima facie* case in favor of the value of deriving indicators from texts and we have concentrated on demonstrating something of the rich variety of ways in which this may be done. This hardly exhausts the topic. Our aim instead has been to open it for further investigation.

References

- [AGI07] N. Archak, A. Ghose, and P. Ipeirotis, *Show me the money! Deriving the pricing power of product features by mining customer reviews*, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007) (San Jose, CA), ACM, August 2007.
- [Bei06] Eric D. Beinhocker, *The origin of wealth: Evolution, complexity, and the radical remaking of economics*, Harvard Business School Press, Boston, MA, 2006.
- [BGI08] K. Balakrishnan, A. Ghose, and P. Ipeirotis, *The impact of information disclosure on stock market returns: The Sarlanes-Oxley Act and the role of media as an information intermediary*, Workshop on Economics and Information Security (WEIS 2008) (Dartmouth College), June 2008, File: <http://weis2008.econinfosec.org/papers/Ghose.pdf>.
- [BK02] David C. Blair and Steven O. Kimbrough, *Exemplary documents: a foundation for information retrieval design*, Information Processing and Management **38** (2002), no. 3, 363–379.
- [Cec05] M. Cecchini, *Quantifying the risk of financial events using kernel methods and information retrieval*, Ph.D. thesis, University of Florida, Gainesville, FL, 2005.
- [CKL04] Gary T. Chen, Steven Kimbrough, and Thomas Lee, *A note on automated support for product application discovery*, Proceedings of the Fourteenth Annual Workshop on Information Technologies and Systems (WITS2004) (Washington, D.C.) (Amitava Dutta and Paulo Goes, eds.), December 2004, pp. 128–133.
- [DKP00] Garrett O. Dworman, Steven O. Kimbrough, and Chuck Patch, *On pattern-directed search of archives and collections*, Journal of the American Society for Information Science **51** (2000), no. 1, 14–23.
- [Dwo99a] Garrett O. Dworman, *Pattern-oriented access to document collections*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1999, Available as a working paper, Department of Operations and Information Management, [Dwo99b].
- [Dwo99b] ———, *Pattern-oriented access to document collections*, Working paper 99-12-20, University of Pennsylvania, Department of Operations and Information Management, Philadelphia, PA, December 1999.
- [FS07] Ronen Feldman and James Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*, Cambridge University Press, Cambridge, UK, 2007.
- [Fuk95] Francis Fukuyama, *Trust*, The Free Press, New York, NY, 1995.
- [GT94] Theodore S. Glickman and Karen S. Terry, *Using the news to develop a worldwide database of hazardous events: A report of the results of a 75-day experiment, with recommendations for further action*, National Science Foundation research grant no. SBR-9309369 report, Center for Risk Management, Resources for the Future, Washington, DC, 30 August 1994.
- [HL04] M. Hu and B. Liu, *Mining and summarizing customer reviews*, Proceedings of Knowledge Discovery in Databases 2004 (KDD04), 2004.
- [JM02] Peter Jackson and Isabelle Moulinier, *Natural language processing for online applications: Text retrieval, extraction and categorization*, John Benjamins Publishing Company, Amsterdam, The Netherlands and Philadelphia, USA, 2002.
- [KMR07] Steven O. Kimbrough, Ian MacMillan, and John Ranieri, *Process and system for matching products and markets*, United States Patent 7,257,568, 14 August 2007, www.uspto.gov.
- [KMRT07] Steven O. Kimbrough, Ian MacMillan, John Ranieri, and James D. Thompson, *Categorized document bases*, United States Patent Application 20070106662, 10 May 2007, www.uspto.gov.
- [Kon06] Manu Konchady, *Text mining application programming*, Charles River Media, Boston, MA, 2006.
- [Lee04] Thomas Y. Lee, *Use-centric mining of customer reviews*, Proceedings of the 2004 Workshop on Information Technology and Systems (WITS), 2004.
- [Lee08] T. Lee, *Learning industry-specific voluntary disclosures from SEC 10-K regulatory filings*, Winter Information Systems Conference (University of Utah, UT), March 2008.
- [LHC05] B. Liu, M. Hu, and J. Cheng, *Opinion observer: Analyzing and comparing opinions on the web*, Proceedings of WWW 2005, 2005.
- [Li06] Feng Li, *Do stock market investors understand the risk sentiment of corporate annual reports?*, Working paper SSRN 898181, University of Michigan, Ann Arbor, MI, 2006.

- [LLP07] Hady W. Lauw, Ec-Peng Lim, and HweeHua Pang, *TUBE (Text-cUBE) for discovering documentary evidence of associations among entities*, Proceedings of the 22nd Annual ACM Symposium on Applied Computing (SAC'07) (Seoul, Korea), ACM, 11–15 March 2007, <http://www.acm.org/conferences/sac/sac2007/>, pp. 824–828.
- [LLW08] T. Lee, S. Li, and R. Wei, *Needs-centric searching and ranking based on customer reviews*, IEEE Conference on Electronic Commerce (Washington, D.C.), IEEE, July 2008.
- [Mie05] Katharine Mieszkowski, *Steal this bookmark!*, Salon, www.salon.com, at <http://dir.salon.com/story/tech/feature/2005/02/08/tagging/index.html>, February 2005.
- [Moe00] Marie-Francine Moens, *Automatic indexing and abstracting of document texts*, Series: The Information Retrieval Series, vol. 6, Springer, Germany, 2000, ISBN: 978-0-7923-7793-1.
- [NY03] Tetsuya Nasukawa and Jeonghee Yi, *Sentiment analysis: Capturing favorability using natural language processing*, Proceedings of the Second International Conference on Knowledge Capture (K-CAP 2003), October 2003.
- [PE05] A.-M. Popescu and O. Etzioni, *Extracting product features and opinions from reviews*, Proceedings of HLT-EMNLP 2005, 2005.
- [Put00] Robert D. Putnam, *Bowling alone: The collapse and revival of American community*, Simon & Schuster, New York, NY, 2000.
- [SBC⁺07] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and J. Chun, *Red Opal: Product-feature scoring from reviews*, ACM Conference on Electronic Commerce (San Diego, CA), ACM, June 2007.
- [SWY75] G. Salton, A. Wong, and C. S. Yang, *A vector space model for automatic indexing*, Communications of the ACM **18** (1975), no. 11, 613–620.
- [Vos07] Jakob Voss, *Tagging, folksonomy & co – renaissance of manual indexing?*, Proceedings of the International Symposium of Information Science, 2007, pp. 234–254.
- [YNBN03] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack, *Sentiment analyzer: Extracting of sentiments towards a given topic using NLP techniques*, The Third IEEE International Conference on Data Mining (ICDM '03), November 2003.