Exploring the New York Times Corpus with NewsClub

Christian Kohlschütter
L3S Research Center / Leibniz Universität Hannover
Appelstr. 9a, 30167 Hannover
Germany
kohlschuetter@L3S.de

ABSTRACT

In this HCIR 2010 challenge paper, I report on the evaluation of the NewsClub information retrieval system on the New York Times corpus.

1. INTRODUCTION

Keyword queries are still the easiest and common way to start a search in an information retrieval system. However, in many cases these queries can be over- or underspecified, and so the quality of the returned results is strongly dependent on the quality of the specified terms. As a complementary feature to unstructured text queries ("is there a document containing the keyword K?"), structured classifications (complex taxonomies or orthogonal, faceted categories) can help the user to choose the right document features to quickly drill-down to a particular aspect within the whole result set, such that only specifically relevant documents are returned. Additionally, the proposed classifications themselves may satisfy the user's information need, especially when looking for aggregate information or when just checking for the presence of a particular classification ("which *authors* wrote the most articles concerning X?", "is there a *book* about Y written by author Z in *2010*?").

However, maintaining a structured classification system requires a substantial, continuous effort, deep domain knowledge and preferably control over the document creation process to ensure a somewhat complete and valid classification, as incomplete or wrong classifications may drastically deteriorate the search experience (especially for the aggregate queries described above). In any case, especially with growing collections such as in the news domain, it will in any case be no perfect or concluded labeling.

In this paper, I present my information retrieval system *NewsClub*, which provides, in addition to keyword queries and faceted search, a third way to explore the set of retrievable documents: relevant terms and phrases. For example, when querying for "terrorism", the system determines *Al Qaeda* and *Osama bin Laden* as highly relevant terms. These n-grams are relatively easy to determine from any unstructured text, do not need any manual processing and yet provide a surprisingly high utility for search. In combination with keyword and faceted search, this allows a very quick and efficient navigation within the result set.

In addition to just showing the most relevant terms and phrases for a query, NewsClub can also visualize term associations (i.e., determine words that deal with different aspects of terrorism, for example: *Afghanistan/Taliban* vs. *Palestina/Hamas*) and find contrasting terms which best match individual sub-queries, i.e.., terrorism w.r.t. Israel/Palestina (*Gaza strip, Intifada*), Afghanistan (*bin Laden*, *Taliban*), Iraq (*Saddam, Zarqawi*) or the USA (*Oklahoma*, *Littleton*). The sub-queries may represent any other typical keyword or classification feature and thus are easy to specify.

While the NewsClub platform has initially been targeted mainly at analyzing news, it is also being deployed in different scenarios. A publicly available system is *NewsClub im Bundestag*[1], which monitors the plenary sessions in the German Parliament. It allows to drill-down by speaker, party, role and legislative period and provides the very same text-analytics features described above.

2. USER INTERFACE

UI Fundamentals. NewsClub's Web-based user interface has been designed for the curious searcher who wants to interact with the system, but it also allows for fast ad-hoc queries. The AJAX application has been built using the Google Web Toolkit, designed for low-latency and high extensibility. The fundamental UI structure was inspired by the Eclipse platform, which allows to have different tabbed panes in one perspective; NewsClub's search perspective consists of a keyword query box and a set of tabs, each containing a different view on the query or the results. The tabs are grouped into the two parts of a horizontally-split pane, a wide pane on the left, and a tall pane on the right, thus allowing the user to activate two different tabs at once, a large one and a small one. Currently, the user can choose from the following tabs. *Left*: Search results, Associator, Contrastor, Time Window and Detail view. *Right*: Facets, Term Stack, Sub-queries (see Figure 4).

Search results. In this view, individual search results are shown (very much like in traditional search engines). The only notable difference is that NewsClub's view supports continous scrolling (one can actually scroll through all retrievable results). Clicking on a result will open the original URL in another window or the locally stored information in the **Detail** view. The **Facets** view contains a tree panel containing all possible facet dimensions, the categories and the corresponding number of matching documents for the current query. By clicking on a category, one can narrow the search to only those documents that are labeled with the

Copyright is held by the author/owner(s).
ACM X-XXXXX-XX-X/XX/XX.

[1] `http://newsclub.de/bundestag`

selected category (the query box is extended by an appropriate button to indicate the drill-down). The **Term Stack** view contains a set of relevant terms and phrases that are matched in the documents retrieved for the query. The relevance of a term is determined by a measure based on the Shannon entropy, not by absolute frequency. The *stack* differs from a typical *cloud* in the fact that the terms are first grouped by importance (only changes in the order of magnitude are considered), then alphabetically. This allows a much faster reception of the most important terms, as they will always appear first (and larger than the other terms). **Associator.** For ambiguous queries, this view helps clustering the relevant terms and phrases into distinct subgraphs. The graph can be zoomed and the terms can be added to the keyword search by clicking them. Using the **Time Window** view one can narrow the search to items of a particular time span. The view contains calendars for start and end date as well as a graph depicting the absolute number of items per date. The values can be smoothed using an arbitrary sliding average and can thus also be used for a rough trend analysis (also see Figure 1). In the **Sub-queries** view the user may specify additional queries that are evaluated with respect to the main query. Generally, sub-queries may be utilized in any other view, but currently, this is only used for the **Contrastor** view. Here, the user can re-arrange the top-500 relevant terms and phrases with respect to relevance to the main query, to the subset of results that also match each sub-query as well as to the variance between the individual sub-queries.

3. EVALUATION ON THE NYT CORPUS

In this section, I report on how one can conduct searches in NewsClub for the task scenarios described in the HCIR2010 challenge call.

Subway crime. We start by selecting "Crime and Criminals" in the "Descriptors" facet, "New York City" at "Locations" and enter "subway OR metro" as the keyword query. We switch to the sub-queries tab, enter the article years 1987 to 2007 and then open the *Contrastor* view. The unfiltered top-10 terms for each year are shown in Table 1. From these terms, we see for example that there was a shooting involving a person called "goetz" in 1987, which we can confirm by adding "goetz" as an additional keyword, switching to the search results view and examining the 27 matching articles (Bernhard H. Goetz was accused for shooting four black youths on a subway train). As another example, in 1990, Brian Watkins, a tourist from Utah, has been murdered and robbed in the subway (first, second and eigth term matched). We may also drill deeper into each year's terms by switching to the *Associator* view. Here, for example, we find out that in 1987 there are many terms around the Goetz case, but that there also was also homeless man being pushed in front of a subway and that a subway token booth has been attacked and one clerk has been shot and critically wounded.

Development of pizza prices. NewsClub currently does not analyze numbers, so this is out of scope at the moment. However, we can search for "pizza" and analyze the Associator graphs (see for example Figure 2). Apparently, some restaurants are famous, some had to be closed due to bad hygienic conditions, and there also is a Mafia connection.

Rent Control. We start a new search for *rent control* with facet "Location:New York" and the sub-queries "pro", "against" and "pros and cons". For *pro*, we find terms like *fund hotel, pro-landlord* and *cloward and piven.* The last term refers to two authors of a study, pro-landlord leads us (via the Associator) to the related terms "pro tenant" and "rent control regulations", and finally "fund hotel" refers to the phrase "hedge fund hotel" (finding one article, where *rent* and *control* are not referring to a phrase). For the sub-query *against*, we find terms like "luxury decontrol" and "vacancy decontrol", and for the sub-query *"pros and cons"* we get "landlord-tenant", "Lansco" (a brokerage firm specializing in retailing properties, according to a snippet summary in the search view), "bedroom apartments", "one-bedroom", and "commercial space".

Adding one or more of these terms to the keyword query finally yields few, highly relevant documents. For example, "rent control 'pro-landlord' 'pro-tenant"' yields 3 results, including "*A Landlord's Lot Is Sometimes Not an Easy One*" and *"Raising the Rent, and Raising the Roof"*, which provide some of the desired arguments pro and against rent control.

Clinton Impeachment. We could start our search the same way like for *Rent Control*, but this time we look at the Term Stack (Figure 3). Without knowing the background of the impeachment, we see that Monica Lewinsky and Paula Jones were involved in a Clinton scandal, that the president lied and that there was the question of "crime or misdemeanor". We also find other persons involved, e.g,. Kenneth Starr and other scandals such as "filegate" and "travelgate", which we may use to further explore the result set by adding one or more of these terms to the query.

Free Concerts in New York City. Let us try searching for "concert ('free admission' OR 'no admission charge' OR 'admission: free')" and narrow to facet Location: New York City. Of the 25 documents we found, 15 have been annotated for the "Organizations" facet, which we can simply open and read the matching terms (see Table 2). We would now have to check each organization by inspecting the search results to determine whether they really offer free concerts.

Member of the Communist party. This task requires a join over the *People* facet, between search results for "Communist party" and "(legislative OR executive)" (both restricted to "Location: New York State"). Matching people are likely to answer the question (but still need to be reviewed). NewsClub currently does not support joins, but luckily, we only get eight labels in "people" for the query "Communist party", which we can manually join with the second query (see Table 3). We can check the remaining five people by again switching to the "Communist party" results (three documents only!). For Governor Mario Cuomo, we see that he only visited a member of the communist party on a diplomatic mission, so we can exclude him from the candidate list. On further reading, we find an article ("*Cardinal and Mayor: Excerpts From Book*") containing a book excerpt by Edward Koch who stated that there indeed was a New York legislator who was an active member of the Communist party and sat on the executive board of a local Communist organization (Koch probably knows what the name of the mentioned legislator is; I do not know).

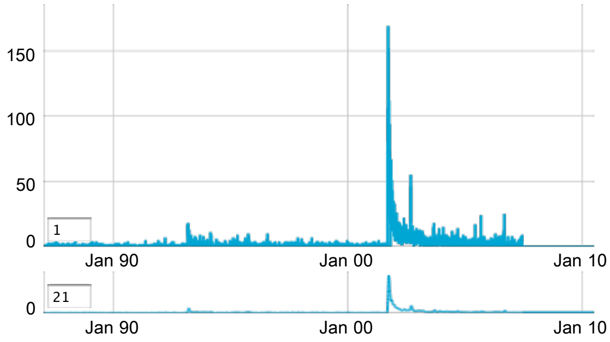


Figure 1: Time Window: "*World Trade Center*"

4. REMARKS, CONCLUSIONS AND FURTHER WORK

The HCIR 2010 challenge is a great opportunity to evaluate the NewsClub information retrieval system on the New York Times corpus. I was able to show that NewsClub can deliver (at least approximate) results for all demanded tasks, without much preparation. The data was loaded into the system basically without any post-processing, reasoning on facet dimensions etc. Indeed, there are some problems with the NYT classification that should be fixed for a production system. For example, there is a change from all-uppercase names to regular case in 1996, which would require special alignment to unify pre- and post-1996 labels. The classification also lacked proper aggregate information/meronym-relations, which would have helped in the task scenarios (for example, it would be good to know that New York City is part of New York State).

On the other hand, also the terms that NewsClub automatically determined relevant appear unfiltered and sometimes cannot be understood without specific knowledge (e.g., person names). A pre-processing using a part-of-speech tagger or a thesaurus would have been helpful.

Moreover, the subway crime task has shown that if we really only want to evaluate specific terms (in our case: a finite set of subway crime words), we would need the ability to specify these terms in the Contrastor. This features is currently being developed, but was unfortunately not ready for this report.

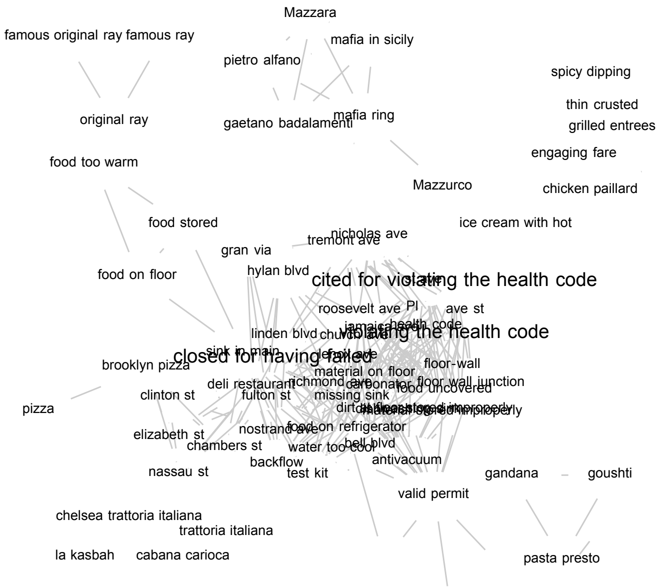


Figure 2: Associator graph for "Pizza" in New York from 1987 to 1996

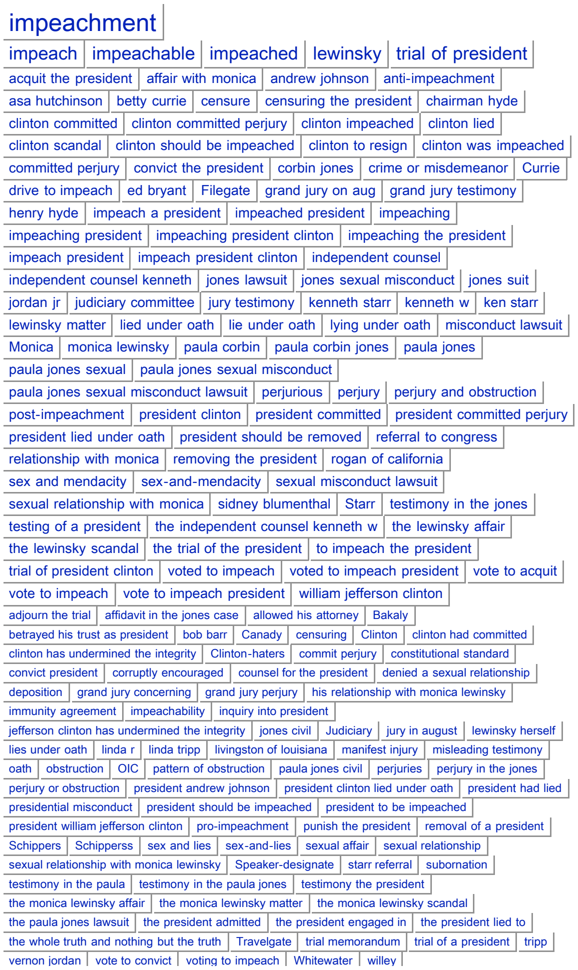

impeachment
impeach | impeachable | impeached | lewinsky | trial of president
acquit the president | affair with monica | andrew johnson | anti-impeachment
asa hutchinson | betty currie | censure | censuring the president | chairman hyde
clinton committed | clinton committed perjury | clinton impeached | clinton lied
clinton scandal | clinton should be impeached | clinton to resign | clinton was impeached
committed perjury | convict the president | corbin jones | crime or misdemeanor | Currie
drive to impeach | ed bryant | Filegate | grand jury on aug | grand jury testimony
henry hyde | impeach a president | impeached president | impeaching
impeaching president | impeaching president clinton | impeaching the president
impeach president | impeach president clinton | independent counsel
independent counsel kenneth | jones lawsuit | jones sexual misconduct | jones suit
jordan jr | judiciary committee | jury testimony | kenneth starr | kenneth w | ken starr
lewinsky matter | lied under oath | lie under oath | lying under oath | misconduct lawsuit
Monica | monica lewinsky | paula corbin | paula corbin jones | paula jones
paula jones sexual | paula jones sexual misconduct
paula jones sexual misconduct lawsuit | perjurious | perjury | perjury and obstruction
post-impeachment | president clinton | president committed | president committed perjury
president lied under oath | president should be removed | referral to congress
relationship with monica | removing the president | rogan of california
sex and mendacity | sex-and-mendacity | sexual misconduct lawsuit
sexual relationship with monica | sidney blumenthal | Starr | testimony in the jones
testing of a president | the independent counsel kenneth w | the lewinsky affair
the lewinsky scandal | the trial of the president | to impeach the president
trial of president clinton | voted to impeach | voted to impeach president | vote to acquit
vote to impeach | vote to impeach president | william jefferson clinton
adjourn the trial | affidavit in the jones case | allowed his attorney | Bakaly
betrayed his trust as president | bob barr | Canady | censuring | Clinton | clinton had committed
clinton has undermined the integrity | Clinton-haters | commit perjury | constitutional standard
convict president | corruptly encouraged | counsel for the president | denied a sexual relationship
deposition | grand jury concerning | grand jury perjury | his relationship with monica lewinsky
immunity agreement | impeachability | inquiry into president
jefferson clinton has undermined the integrity | jones civil | Judiciary | jury in august | lewinsky herself
lies under oath | linda r | linda tripp | livingston of louisiana | manifest injury | misleading testimony
oath | obstruction | OIC | pattern of obstruction | paula jones civil | perjuries | perjury in the jones
perjury or obstruction | president andrew johnson | president clinton lied under oath | president had lied
presidential misconduct | president should be impeached | president to be impeached
president william jefferson clinton | pro-impeachment | punish the president | removal of a president
Schippers | Schipperss | sex and lies | sex-and-lies | sexual affair | sexual relationship
sexual relationship with monica lewinsky | Speaker-designate | starr referral | subornation
testimony in the paula | testimony in the paula jones | testimony the president
the monica lewinsky affair | the monica lewinsky matter | the monica lewinsky scandal
the paula jones lawsuit | the president admitted | the president engaged in | the president lied to
the whole truth and nothing but the truth | Travelgate | trial memorandum | trial of a president | tripp
vernon jordan | vote to convict | voting to impeach | Whitewater | willey

Figure 3: Term Stack for "*Impeachment Clinton*"

Table 1: *Subway crime.* Top-10 words and phrases for years 1987 to 2007.

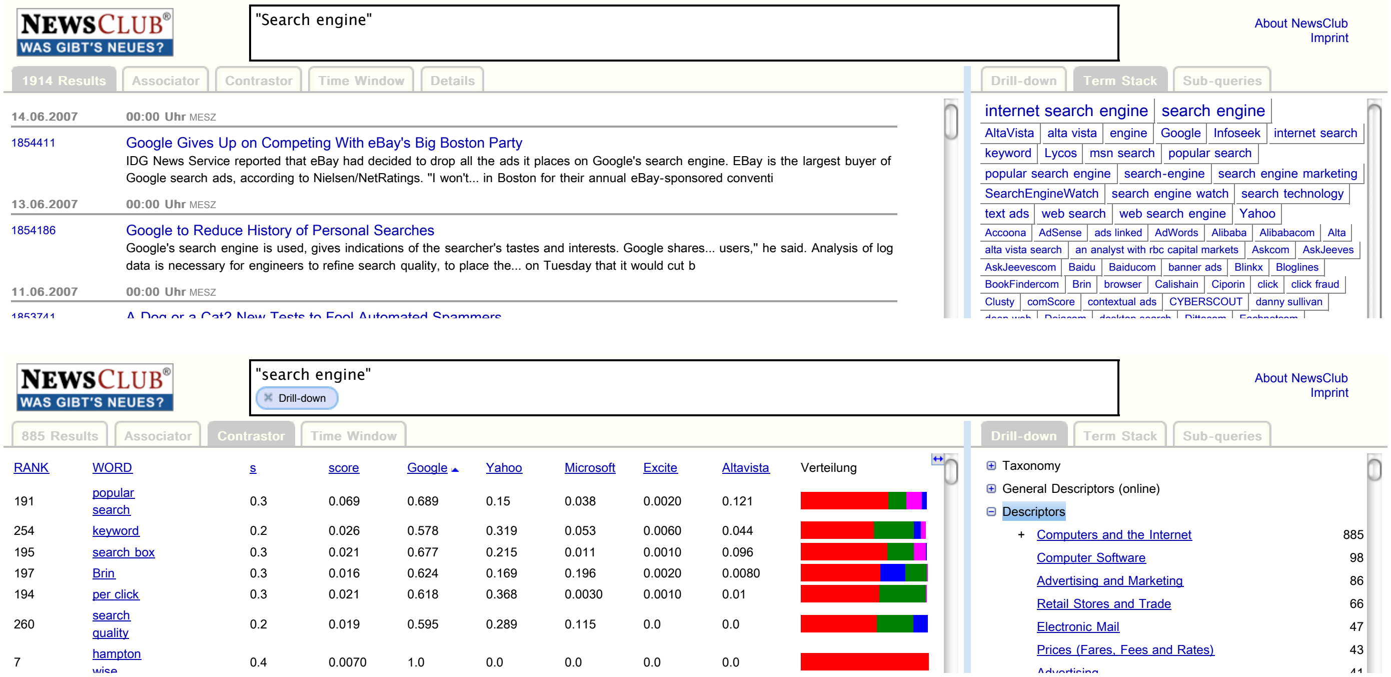
1987: DelCastillo, goetz shot, crime at kennedy, disrupted subway, woman eight, bernhard h, berhard goetz, professor kaufman, downtown irt, goetz verdict. **1988:** vincent del, albert o, bmt station, del castillo, transit patrolmen, debra elisa, Moraff, rider-advocacy, stranger-to-stranger, fatally burned. **1989:** sergeant galea, miss honig, reported in new york city rose, Decepticons, jackie peterson, sergeant keaveny, Nero-like, abandonned, citing a sharp, neck-bending. **1990:** tourist from utah, brian watkins, Gosso, men commit, larcenies, arrested teen, civilianized, utah tourist, non-negligent, anti-crime. **1991:** token-booth, token booth clerk, pickpocketing, subway tracking, thomas reppetto, toughest mayor on crime, Debhasis, Dettman, Onionhead, Rettler. **1992:** fare-evasion, Cantius, Taneka, Gasparik, Pecola, stage-prop, chain snatching, eexit gate, Tirsa, abuna paulos. **1993:** ted husted, Ficaro, Kowslowsky, Unick, gerry griffin, nostaglic, co owner of la, convering, jamican drug, kevin jett. **1994:** Del-Debbio, peter del debbio, desmond robinson, Darnal, shaul linyear, peter del, hate to hate, Coplen, robbery on feb, allyn winslow. **1995:** appeared in news, questioned the veracity, Brahmbhatt, phenemenon, Lanzman, Maioglio, pushed in front of a subway train, Bonina, joseph castellano, copulated. **1996:** violent and property, Elmer-DeWitt, Maxian, the merger of the transit, single-officer, car-window, solo patrol, strongest democratic, two-officer, murder of brian. **1997:** larcenies, transit division, convicted of the misdemeanor, Vimala, armed teen, survey-research, overall crime, Ceasia, mood-shifting, sergeant miranda said. **1998:** bap bap, captain, phipps, th street and roosevelt, murder or rape, Kolden, Petracco, cab watch, DeMarion, crime has dropped. **1999:** misdaemeanor criminal, Ciraolo, bat as a weapon, Lombardino, fare gates, throat-grabbing, fare beating, danied that the department, fordham students, domestic-relations. **2000:** max fine, Haly, subject to sexual, credit-taking, struggling-artist, murder-conspiracy, Kelling, Jaycor, older-brother, professor at the university of california in los angeles. **2001:** alibi statement, pre-arraignment, wwwnytimescom/metro, open-aired, spokeswoman for the new york city law department, Calik, Gulsen, Huascar, resulting emotional, multiborough. **2002**: murder nine, John/Jane, half-green, Eksi, st precinct station, leave-me-alone, drug-drenched, stuff of hollywood, unusually loud, exxon gas. **2003:** nd street and seventh, william glaberson, trial of peter, deserted station, Cassarino, assemblyman ivan, peter gotti, frederic block, DeFede, crime-reduction. **2004:** captain matusiak, Facciolo, Wolfrom, assassination-style, Stacy-Ann, pimple-facet, twice-broken, Fanale, dead-on-arrive, phone for hours, rape and beating of a jogger in central. **2005:** Kneafsey, contact with the homeless, metrocard vending, living in the subway, chambers street subway station, digitalize, trust necessary, theft-deterrent, Wastberg, madrid train bombing. **2006:** jersey trucking, courtroom space, citing continuing, lag times, sentenced yesterday in federal district, arrest-to-arrignment, councilman peter, spokesman for mayor michael r, turnstiles. **2007:** mappelle, Lucyna, producing the latest, palazzolike, security is concerned, crime and vandalism, behind the crime, Karnen, Eterno, urinators.

Table 2: *Free Concerts.* Retrieved organizations from NYT facet classification.

92d Street Y, Brooklyn Children's Museum, Fogg Art Museum (Cambridge, Mass), Gardner, Isabella Stewart, Museum (Boston), Halle Orchestra, Liberty Science Center (Jersey City, NJ), Long Island Rail Road Co, Lower Manhattan Development Corporation (NYC), Museum of Fine Arts (Boston), Museum of the City of New York, NYC-TV (Cable Station), New York Botanical Garden, Newark Museum (NJ), Queens Wildlife Center, World Trade Center Memorial Foundation

Table 3: *Communist Party/Legislative or executive post*: People

"Communist Party": *Cuomo, Mario M (Gov); Dionne, E J Jr.; Jackson, Jesse L (Rev); Kazakov, Vasily (Deputy Chmn); Koch, Edward I (Mayor); Mailer, Norman; Schmalz, Jeffrey; Vinogradov, Vladimir M (Min)*
"Legislative OR executive" *Cuomo, Mario M (Gov); Dionne, E J Jr.; Jackson, Jesse L (Rev); Koch, Edward I (Mayor); Schmalz, Jeffrey*



"Search engine"

14.06.2007 00:00 Uhr MESZ

1854411 Google Gives Up on Competing With eBay's Big Boston Party
IDG News Service reported that eBay had decided to drop all the ads it places on Google's search engine. EBay is the largest buyer of Google search ads, according to Nielsen/NetRatings. "I won't... in Boston for their annual eBay-sponsored conventi

13.06.2007 00:00 Uhr MESZ

1854186 Google to Reduce History of Personal Searches
Google's search engine is used, gives indications of the searcher's tastes and interests. Google shares... users," he said. Analysis of log data is necessary for engineers to refine search quality, to place the... on Tuesday that it would cut b

11.06.2007 00:00 Uhr MESZ

1853741 A Dog or a Cat? New Tests to Fool Automated Spammers

internet search engine | search engine
AltaVista | alta vista | engine | Google | Infoseek | internet search
keyword | Lycos | msn search | popular search
popular search engine | search-engine | search engine marketing
SearchEngineWatch | search engine watch | search technology
text ads | web search | web search engine | Yahoo
Accoona | AdSense | ads linked | AdWords | Alibaba | Alibabacom | Alta
alta vista search | an analyst with rbc capital markets | Askcom | AskJeeves
AskJeevescom | Baidu | Baiducom | banner ads | Blinkx | Bloglines
BookFindercom | Brin | browser | Calishain | Ciporin | click | click fraud
Clusty | comScore | contextual ads | CYBERSCOUT | danny sullivan

"search engine"

RANK	WORD	s	score	Google ▲	Yahoo	Microsoft	Excite	Altavista	Verteilung
191	popular search	0.3	0.069	0.689	0.15	0.038	0.0020	0.121	🟥🟩🟪🟦
254	keyword	0.2	0.026	0.578	0.319	0.053	0.0060	0.044	🟥🟩🟦🟪
195	search box	0.3	0.021	0.677	0.215	0.011	0.0010	0.096	🟥🟩🟪
197	Brin	0.3	0.016	0.624	0.169	0.196	0.0020	0.0080	🟥🟦🟩
194	per click	0.3	0.021	0.618	0.368	0.0030	0.0010	0.01	🟥🟩
260	search quality	0.2	0.019	0.595	0.289	0.115	0.0	0.0	🟥🟩🟦
7	hampton wise	0.4	0.0070	1.0	0.0	0.0	0.0	0.0	🟥

⊞ Taxonomy
⊞ General Descriptors (online)
⊟ Descriptors

+	Computers and the Internet	885
	Computer Software	98
	Advertising and Marketing	86
	Retail Stores and Trade	66
	Electronic Mail	47
	Prices (Fares, Fees and Rates)	43

Figure 4: NewsClub's search perspective. Showing the *Search Results* and *Term Stack* tabs (top) and the *Contrastor and Drill-down* Tabs, restricted to "Descriptors: Computers and the Internet (bottom)