Question #1) How do we make sure we are the leader in search in 2007? What are the new innovations in search, new algorithms, new content? Will universal search work? How do we deal with the problem of "proliferating verticals?" How will we index and handle all the personal and user generated content that is so hard to rank? What information do we not have in our search index and how will we get it?

Bill Brougher, Ben Gomes, Sepandar Kamvar, Marissa Mayer, Udi Manber, Bindu Reddy, Amit Singhal, Johanna Wright

Summary

We believe that Google will remain the leader in search in 2007. This year, however, we must take steps to strengthen our position in the years to come. Specifically, we must:

- Focus on core search and continue to allocate 35% (1/2 of 70%) of engineering new hires to search in order to keep momentum.
- Create a mechanism that provides for aided exploration for users that have trouble formulating their queries.
- Commit to reinvention -- accept that we will make some user experience mistakes but that mistakes are important as they indicate a lack of fear of change and, hence, are a sign of progress and prevent staleness.
- Build on growing momentum around universal search through increased staffing and tactical execution.
- Invest in ranking excellence of a two-tiered index where '[redacted]is the large volume low quality tier.
- Doing different "operations" on results seems very compelling (see on a map, sort by price) but requires structure, the technology to extract it for web content, and an infrastructure to support it. This is a hard problem and we are not focused on it or organized around it, particularly in the case of extraction. We need to organize and staff this effort in 2007 such that we either achieve success or understand with reasonable certainty that it is not feasible in the short term.
- Identify key vertical growth areas and build strength into our core search to support these functions -- identified as particularly important are people, products, and local.
- On the whole, we feel the growth of sites like MySpace and YouTube puts the search business in jeopardy. To secure a safe position in that environment, we should strive to be the search medium for those areas (video, social nets) and also provide some community interaction and entertainment within our core products as well (CCC and iGoogle).
- Turn having the largest user base into an unfair advantage by building out technology that improves linearly with user base size -- initiatives that have this property include: using toolbar data for personalized search, integrating recommendations, question/answering, and creator/trust rank.
- Identify local markets where we are likely to fail in search. Our analysis leads to the conclusion that Arabic, Thai, India, Egypt, Israel, Spanish, and Turkish in particular are at-risk markets. Staff people to analyze, correct, and focus on these markets.

Continued investment in search quality

Observation: We have many promising ranking initiatives underway for the coming year. Recent gains indicate there is significantly more possible on core ranking. Initiatives include:

- Continued investment in our core ranking via query and document understanding.
- Continued investment in user signals, like clicks. Our search users create the first level of network effect for search quality and we are investing in this heavily.
- Hard queries: Queries for which users are frustrated even when they have told Google all they could, there is a strong effort to improve user experience for such queries.
- Query structure analysis: identify different types and look at past usage of those queries to improve ranking
- Suggestions for popular queries
- Rankboost: continue developing our learning system to take human rating data as input and predict new ranking signals
- Low link and small corpus ranking: important for integrating new deep web and mom & pop and international content
- Non-web ranking in preparation for universal search: improve and standardize ranking for other properties by applying tried and true web search techniques, augmented with domain information
- Continued work on personalization.

Remedy/Initiative to address: Situation doesn't merit remedy. We have built good momentum here through increased allocation and effort. We should continue to allocation 35% of new engineers to Search Quality (half of 70%).

Observation: Often users do not look for something specific, do not know how to formulate their query, or do not even know what they're looking for. We will also need to provide an exploration mechanism for concepts. Consider the query [proper spacing in documents] we could suggest to the user [line spacing] which will give much better results, and this direction is part of our query refinement effort. In addition, a better, and much more difficult solution, will be to try to really understand topics in both web pages and queries. The Prose project started in this direction with fixed attributes, user contributed labeling, and few selected domains. We will need to extend it web wide.

Remedy/Initiative to address: Add query refinements to popular and hard queries, and experiment with other ways to help users formulate the right query.

A commitment to reinventing ourselves

Observation: The threat of becoming "our father's oldsmobile" is very real and, due to the accelerated timeline of the web, it won't be along the same timelines (i.e. a generation), it will happen in a matter of years. On the whole, we have become too conservative and anti-change. There have been many observations of this trend -- tremendous pushback from Googlers on small changes like the more box, resistance to things like left nav because they seem big but not compelling enough, etc.

Remedy/Initiative to address: In order breakthrough this, we have to be comfortable with reinvention and change for changes sake. To do this we need to set aggressive goals around user experience changes - even if the changes made are ultimately mistakes - the beauty of running a service is that we can revert those changes or iterate to something new quickly.

Steps to make Universal Search successful

Observation: One of the largest points of user confusion or frustration is not knowing what content we have and how to get it.

Observation: With the deployment of librarian and superroot as well as with a strong team staffed, for the first time ever universal search seems realizable.

Remedy/Initiative to address: Don't let the perfect become the enemy of the good.
Much progress has been made in the past year, but for this project to seem like a reality, the team must focus on an initial launch and then iterate. We will place two stakes in the ground to iterate upon. These launches should happen quickly, ideally in Q4 '06, allowing for continued improvements in 2007.

1. Launch a version of the infrastructure project that will enable universal search in the long run: data updated via Librarian to [redacted] and [redacted] layers, merged via a supperroot and support 100% Web QPS. This launch uses conservative ranking and UI in order to get out the door and not disturb our users.

2. Launch UI and ranking changes for news, local, and images integration. Differential UIs for news and local and increased coverage of the images onebox UI will allow users to start interacting with and seeing the benefits of Universal Search. This launch will also enable us to start seeing how users interact with the product changes and help to drive our direction moving forward.

Observation: To improve comprehensiveness and support Universal Search we plan to grow Teragoogle. We will expand the index to include more content and search [redacted] on every query. We will refresh it weekly significantly improving our age distribution and enabling better fall through from the base index when new content is added to base. There are operational changes that will need to occur in our indexing and serving infrastructures to accommodate this. Also, ranking is not as mature and refined in [redacted] as it is in [redacted]. However, Teragoogle can not do several ranking operations that are necessary for maintaining a state-of-the-art ranking system. Therefore, we will need to keep critical corpora in Mustang to maintain our lead in core properties.

Remedy/Initiative to address: Invest in ranking excellence of a two-tiered index. A richer index, Mustang, for advanced ranking operations to maintain our lead in core areas like web search, and a cheaper index, Teragoogle, which does not provide the same quality as a richer index but is less expensive for serving everything.

Deeper understanding of structured data

Observation: To enrich search, we must develop a deeper understanding of structured data and the powerful operations that having structured data can make possible for our users. Operations like attribute search over jobs, real-estate and personals as well as the ability to slice and dice existing data based on location, price and date.

Observation: We have fledgling extraction efforts underway right now, but not enough of a focus to take all of our petabytes of unstructured data and successfully extract structure.

Observation: This is a hard problem. It is not clear that it is solvable. Our current staffing and organization, however, is sparse enough that it is hindering our ability to assess feasibility.

Observation: We need a way to store and manipulate a massive amount of structured data. Currently that infrastructure is thought to be Google [redacted], but [redacted] is not ready for many of the requirements that will be placed on it by the extraction efforts above or the search needs.

Observation: For Google [redacted] to be successful and to surface on Google.com with any frequency, we must focus on quality.

Remedy/Initiative to address: Fund and organize to succeed in data extraction efforts. Given that the vast majority of our data is unstructured to provide great attribute search we need comprehensiveness. We will first assess the feasibility of extracting location, time and price as they will affect about 20-25% of the queries on Google.com. We also need to decide what infrastructure should hold structured data if and when we are capable of extracting it. Finally, we must get search quality engineers staffed on the task of ensuring quality in Google [redacted] and in extracted data. Key issues to address are attribute convergence and normalization.

Improving web search in high-growth areas that are weak on Google today

Observation: When Google search is deficient in an area, we open up the opportunity for vertical search engines to fill the gap. Thus, as we assess where Google search needs to improve, one method for finding weaknesses is to look at what sites particularly in vertical search are growing quickly (with the assumption that this generally doesn't happen if Google is filling that need well). The team developed the following analysis of areas where vertical search engines are growing quickly and cross-correlated it with 3rd party data:

Priority	Area	Rationale	Initiative
1)	Better name search	Explosive growth of MySpace (2B pvs/day)	Use SPDs to acquire content. Pay social networking sites to enable crawling. Develop an ads product whereby people can submit their names to appear on Google

2)	Better product search	High commercial value (10-12% of searches)	Focus on quality and coverage within Google [redacted] and web ranking
3)	Build an understanding of local and real estate into Google	Fast growth in real estate and compelling opportunity to enhance local	Combine user interface. Enable users to generate content where it does not yet exist online
4)	Enhance experience of job related queries on Google	Drive too much traffic to competitors monster and hotjobs	Focus on quality and coverage within Google [redacted] and web ranking
5)	Improve spammed travel queries	High revenue area with poor quality search results	Improve understanding of spam signal and how they affect this area

(See Appendix for supporting data)

Remedy/Initiative to address: Initiatives noted in 4th column above.

An orthogonal threat: websites with social interaction and high entertainment value

Observation: When reviewing the quickly growing websites (MySpace, YouTube), the team developed an opinion that these social networking sites will ultimately represent a threat to our search business as people will spend more time on those sites and ultimately may do most searches from the search boxes available there. They aren't direct competitors, but they may displace us in end-user time tradeoff. This is particularly true of Live.com whose traffic grew over 1000% as Microsoft transitioned MSN Spaces to Live.com. The analogy that we discussed was a library - Google - in the face of the dawn of movie theaters - MySpace, YouTube - where it even seems possible that the movie theaters will build in bookstores or libraries. While it is comforting that libraries still exist, there's no question that movie theaters now get more consumer time and do more volume. Since it seems unlikely that we will be able to convince people to be entertained less, we feel it is important to develop an entertainment strategy.

Remedy/Initiative to address: Our remedy has four prongs prongs. We need to own the search box on the entertainment sites, we need to be the search site where you can find entertainment content, we need to succeed in social networking, and we need build entertainment and social interaction into our search experience.

1. The first prong falls to our syndication team which should target these rapidly growing sites and win all major syndication deals to own the search boxes.
2. Become the search site for this type of content. This means comprehensive indexing of MySpace, YouTube, NetVibes, RSS, and so on. If we succeed in being the primary way to search these vast arrays of content, this will offer us some protection from these fast growing sites as we will likely grow proportionally to them. We need to staff two efforts, aggressively acquiring this content for indexing and improving low-link ranking.

3. The third prong falls to the CCC team where they need to succeed in social networking (Orkut?).
4. We need to integrate the CCC products as well as more serendipity into our search experiences. We believe that iGoogle, Desktop, and OneGoogle are all good first steps.

More users = Better search; Linear/exponential improvements that leverage Google's user base size

Observation: Google has more users than any other search engine by a large margin. If we could find a way to improve search that scaled linearly, or even better exponentially, with the size of the user base, we could harness an unparalleled advantage.

Remedy/Initiative to address: The best way to find such an advantage is to invest in and explore areas where this kind of finding seems possible.

- Use more pervasively the data that we have available to us, including toolbar and Orkut data. We have a huge number of users and a huge amount of data generated every day that could improve search. We continue to worry about privacy, implications, perceptions, coverage, but we've talked about it for years and we should use toolbar data much more pervasively in 2007.
- Build out recommendations based on user data/behavior. Search is really about recommendations. Even some of our non-search-based products are essentially recommendation engines (News, Desktop, iGoogle). If we can harness signals from user behavior into coherent recommendations, we should be able to build an unsurpassed recommendation engine. In order to do this, we will deploy the generic recommendations infrastructure that the personalized search team has built across various product suites.
- Invest in question-answering. Question-answering remains a highly volatile, highly opportunistic space. We anticipate many changes in this industry in the coming months. If question-answering takes off on in the global market (we can note it already has in Asia), the best question-answering network will be the one with the most users participating and the best sense of user reputation. Our user base and network effects leave us well-positioned. The MUSE project extends the traditional question answering model by adding syndication, one of our core strengths. Questions and answers will be posted on relevant search results, but they will be also syndicated to relevant publishers (using an AdSense-like matching technology) and shown all over the web. Publishers will gain relevant interesting content that will attract users, and they will gain traffic generated from us. We will gain access to users who are more likely to be familiar with the questions. Reputation management will be a core part of this effort.
- Develop creator-centric ranking and browsing of user-generated data. For ranking and discovery of user-generated data, we will develop a ranking of *content creators* and will integrate this ranking as a signal into core search.
- Increase the number of signed-in users and the amount of data per signed-in user in order to broaden the impact of the above three initiatives. Currently, our guarantees of transparency, security, and control allow us only to personalize the Google experience for

those users who are signed in. Today, 8% of results pages are from signed-in users. We plan to increase the number of signed in users with both internal and external strategies. Internally, we plan to (a) experiment with a tiered login system and (b) continue to invest in products that will incent people to both sign in for search such as iGoogle and Google Notebook and bring their data online where it is accessible and searchable by them such as Writely and Picasa Web. Externally, we recommend the acquisition of facebook.com, which we believe will double the number of signed-in searches on google.com [2] and allow us to provide a higher quality of search and recommendations due to the extent of personal data available. An interesting case study is the impact of orkut.com on the number of Brazil signed-in searches. In June of this year, 1% of queries on google.com.br came from signed-in users. Today, 19.45% of queries from google.com.br come from signed-in users, as compared to 8% from the next highest domain, google.com, and 5.2% from google.com.in (also a country with high orkut usage).

Which international markets are in trouble?

Observation about content: In Korea the ratio of pages in our index to active internet users is 8:1, whereas in the US the ratio is well over 100:1. This relative lack of content created an opening for Naver to seed the search market with Knowledge Search content. Our analysis shows risk the same thing happening in smaller markets including Arabic markets, India and Thailand.

Remedy/Initiative to address: Address lack of content by building systems to encourage generation of content in small markets – Arabic markets, India, and Thailand. Our search ranking algorithms are not useful unless there is enough content to rank to have a broad interest search engine. We estimate we need about 200 million documents to return quality results, about the size of the Google News archive search index.

Observation about coverage: We lose in markets where we don't achieve comprehensiveness. Better coverage than allowed Baidu to get a foothold in China in 2001-2004. Our crawl analysis indicates that we have a similar lack of coverage in Spanish speaking countries and in Turkish speaking countries (See appendix International Content Analysis).

Remedy/Initiative to address: Expand coverage in Spanish and Turkish.

Observation about language specific changes: Yandex beat us in Russian because they had a better understanding of the language. Reinforcing this theory, we saw a small change focused on improving morphology grow our market share from 15% to 22%. This highlights the importance of excellent understanding of the language and good language tools.

Observation about focus: We risk losing in markets where we aren't paying attention. For example, last year at this time we were #1 in Eqypt. Today we are #2 to Yahoo!. This has happened because in the past year they had people working on Egyptian and we did not. Similar things are happening in Israel.

Remedy/Initiative to address: The Search Experience Specialist Program (SESP) program will identify and employ have native speakers in different countries to send us feedback on Search Quality and UI.

References

[1] Kamvar, Schlosser, Garcia-Molina. The EigenTrust algorithm for reputation Management in peer-to-peer networks (http://scholar.google.com/scholar?hl=en&lr=&cluster=12096326994709710049)

[2] At approximately 300mm pageviews per day (extrapolated from http://www.facebook.com/jobs_engineering.php), facebook.com has almost as many pageviews as google.com results pages (from https://www.corp.google.com/~logs/newreports/google/2006/10/21/gws.html). If 10% of those pageviews result in a query, then we will have doubled the number of logged in page views. Furthermore, a facebook acquisition would allow our own users to search over a large repository of personal information that google currently does not have access to.

Appendices

Appendix: Vertical Analysis

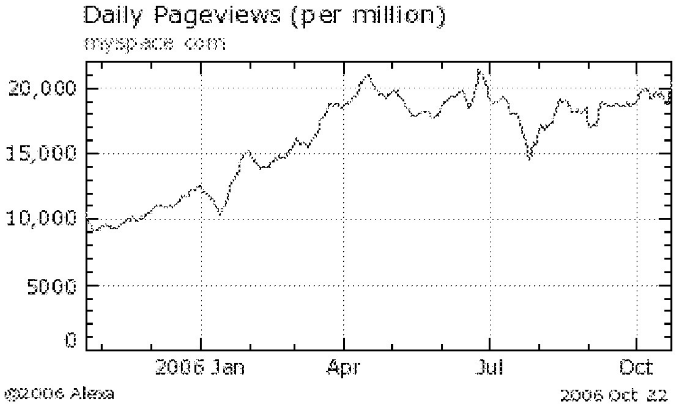
People search and social networking is the fastest growth area on the internet right now with MySpace as the clear leader. MySpace has 2B daily page views and accounts for 11% of Google traffic up from 4% in January, and [myspace] has surpassed [ebay] in the volume of queries issued from google.com. In the past there have been many popular culture references to "googling" someone to find out more. This points to Google's ubiquity and the idea if people do not know where to go to find information, they should go to Google. In the future, we do not want users "myspace" people for more information. Google should host all information about a person, including myspace info.

Product search also represents an area of concern. While we have a high percent of google.com query volume, 10-12% and the most comprehensive data-set (250 M products) for products, we are sorely lagging behind on product search. eBay and Amazon are the leaders in this category with more than 400 M/searches per day. Become.com and eDeals are fast growers in this category quickly surpassing Froogle (Forbes - add to sources). Looking at category growth numbers from comscore, retail remains one of the largest categories.

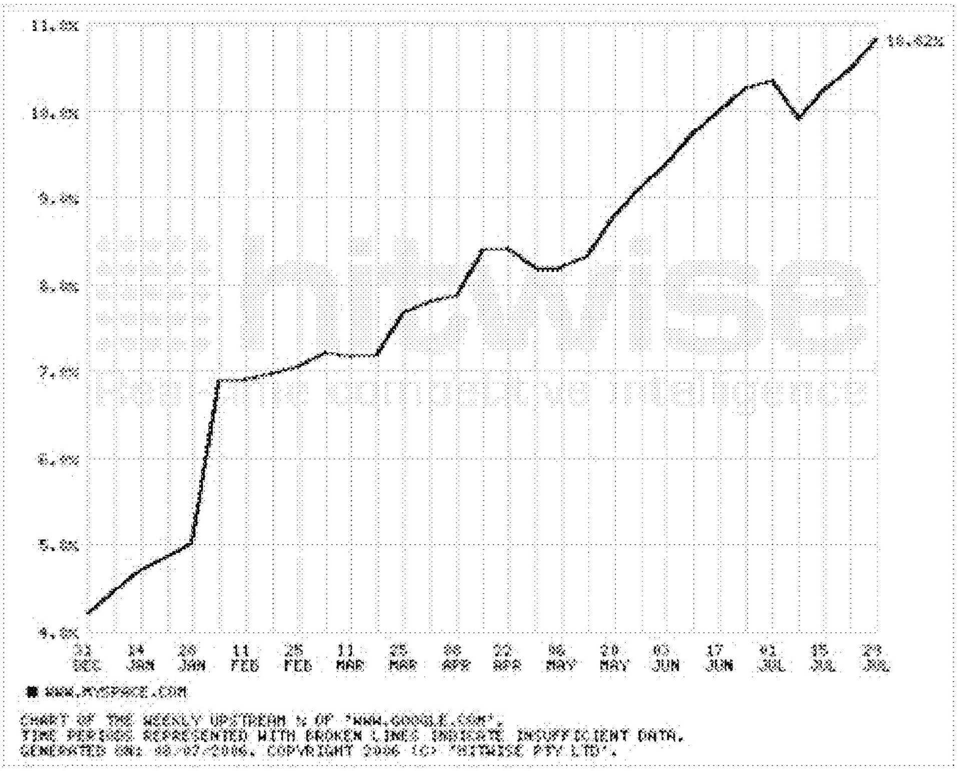
Real Estate is a fast growing vertical and there is no clear winner in this space as yet but Century Remax and Century21. C21 grew 454% in page views but only 1% in unique visitors from Mar-Sept 2006, Remax grew 270% in PV, 2% in UV. So both sites were high growth in that people were looking at a lot of pages -- probably folks looking to purchase RE before interest rates rose any further or people looking to sell before the predicted downturn -- but not a lot of new eyeballs (Fiona Lee's report & ComsScore data)

Myspace page view growth reaches 2B

1 year chart from Alexa Internet shows MySpace reached 2 billion page views in October 2006.



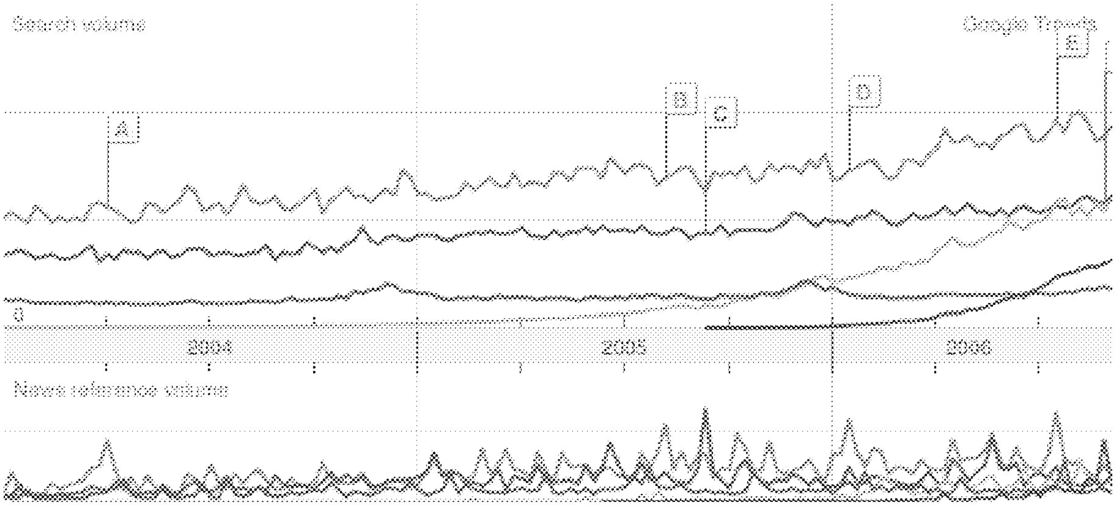
Almost 11% of traffic to Google.com was previously at MySpace



Social networking traffic growth via Google Trends

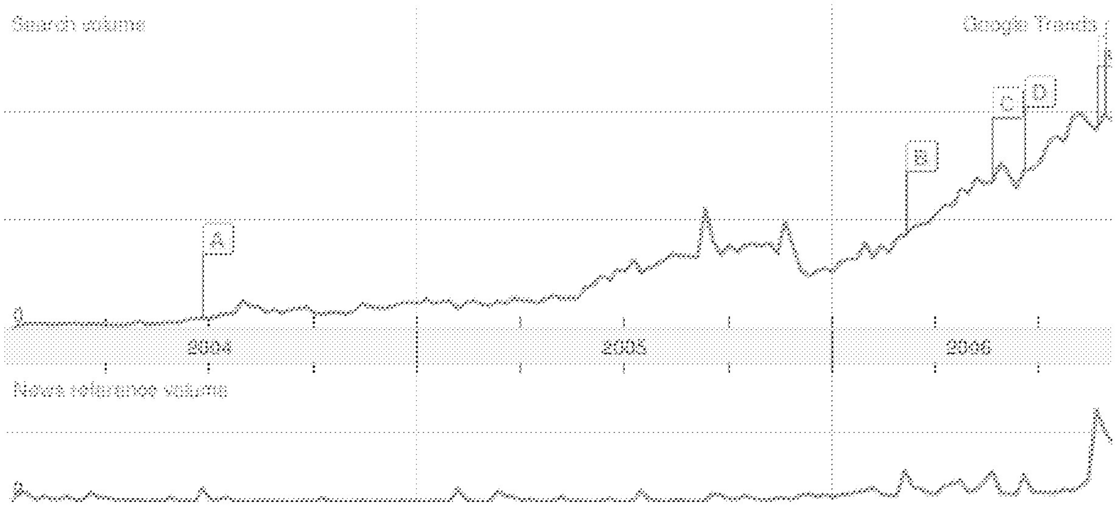
Query volume to [myspace] and [youtube] surpassing popular sites like ebay and amazon.

yahoo ●ebay ●myspace ●amazon ●youtube

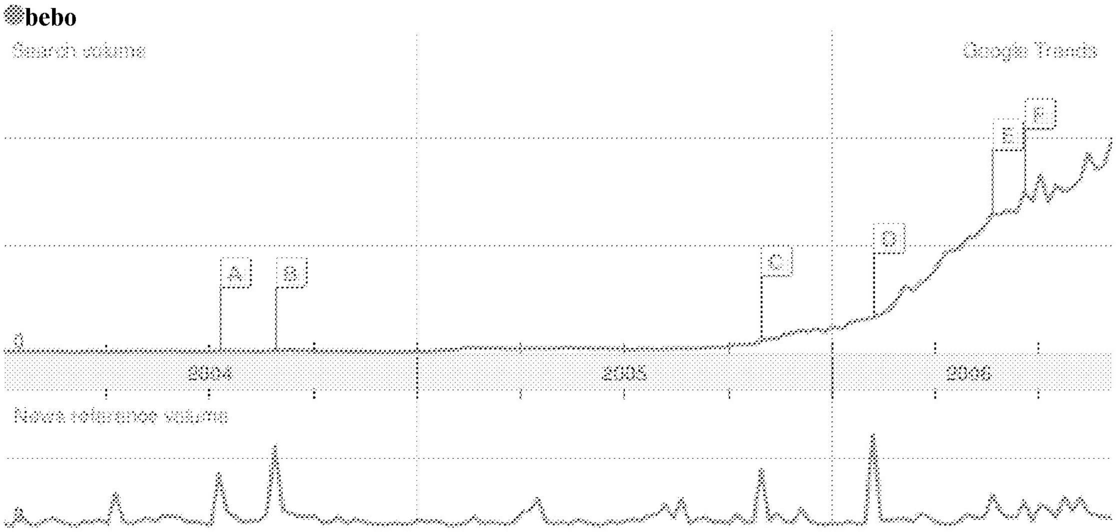


Internationally popular communities sites follow this growth curve, examples are [orkut] and [bebo]

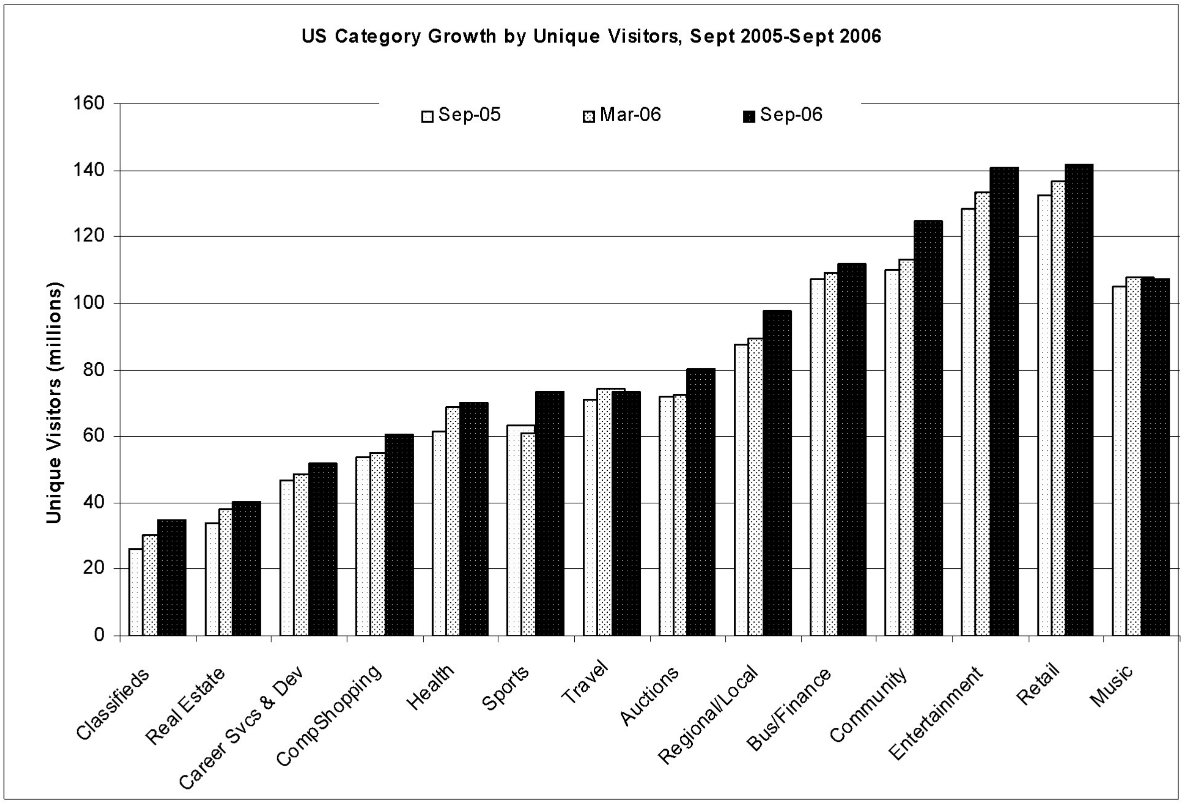
orkut

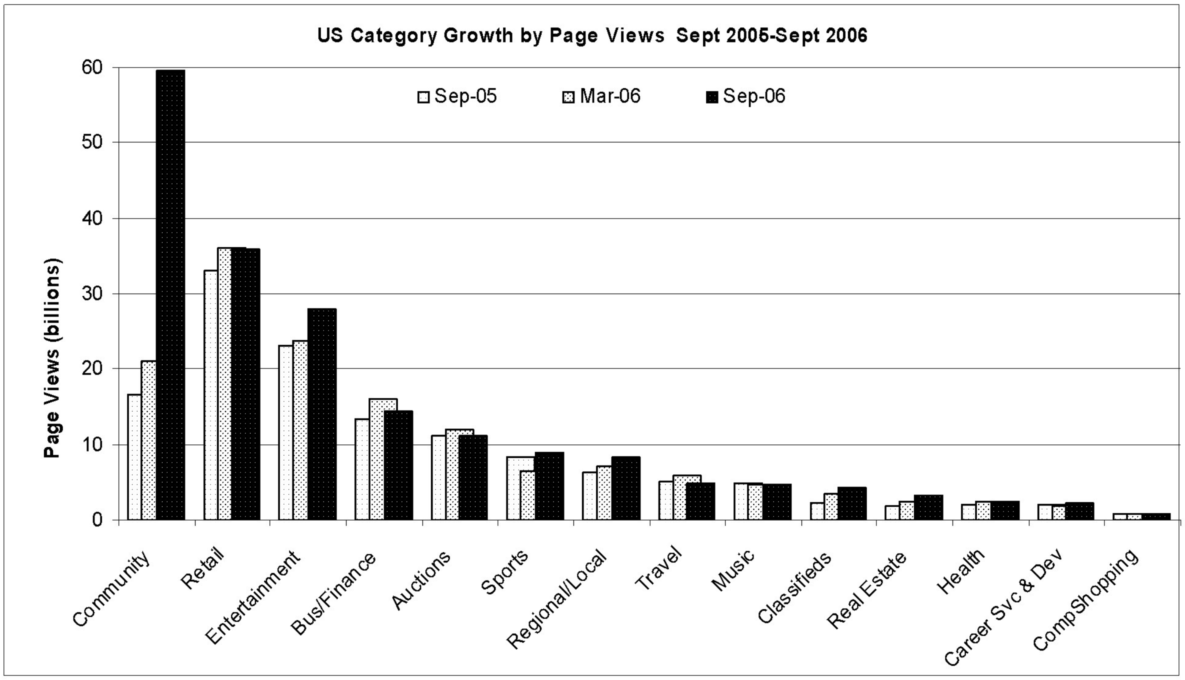


Example 2 [bebo]

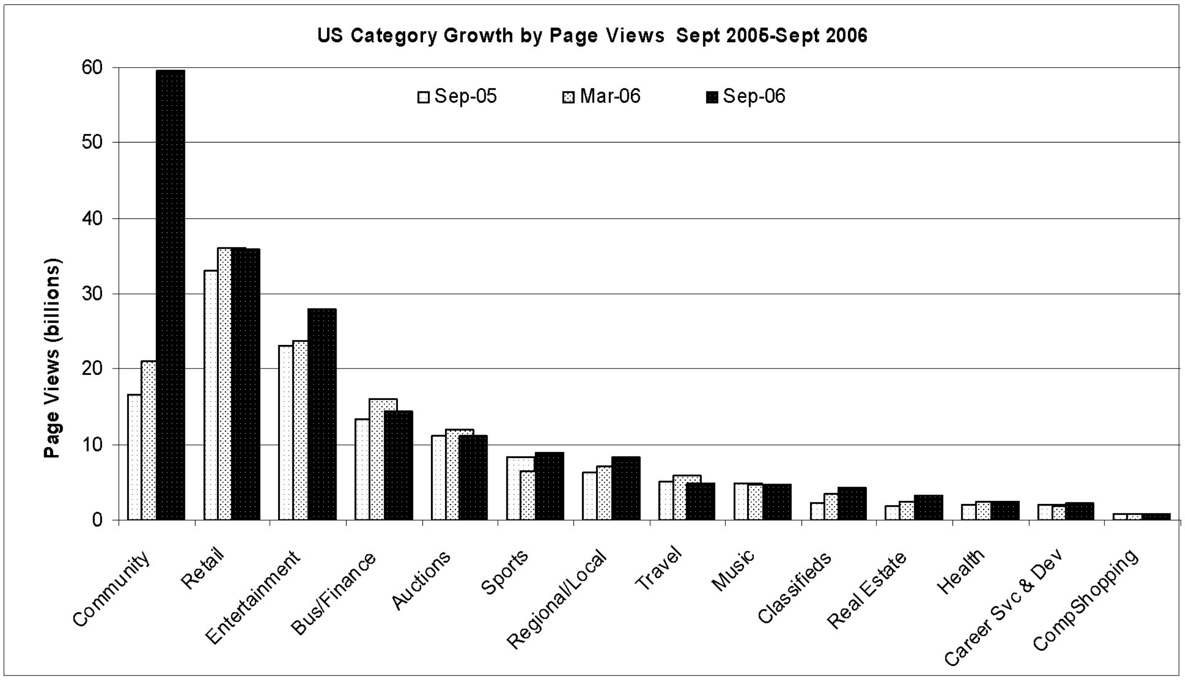


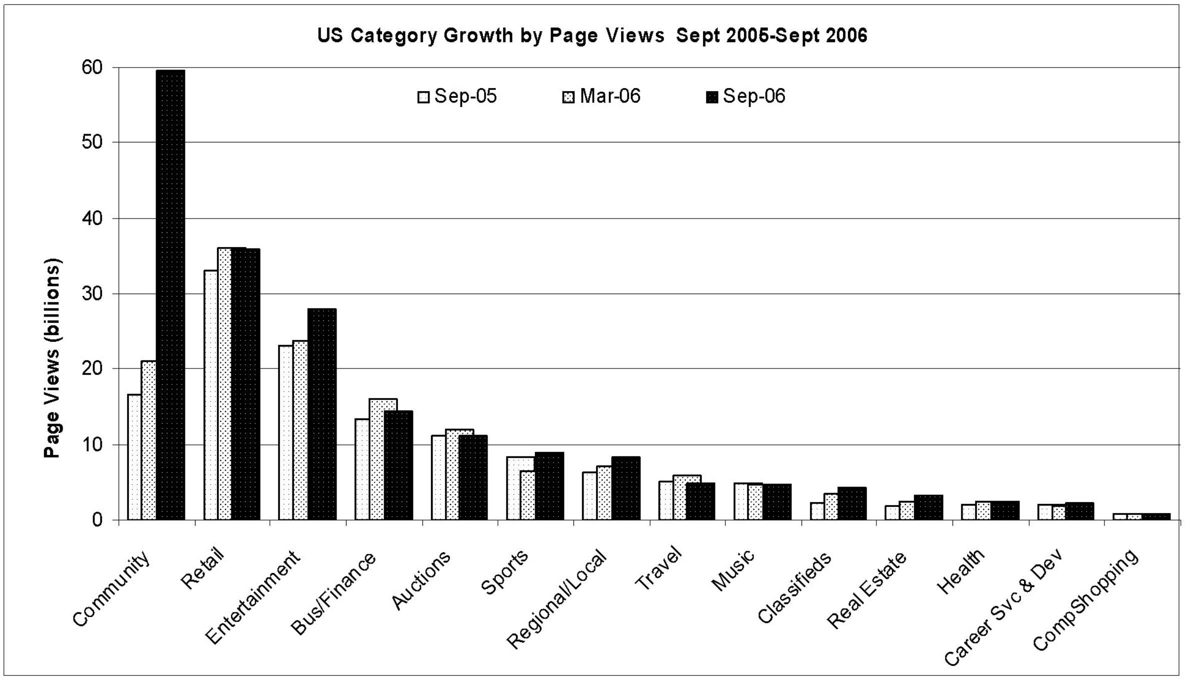
Entertainment, Real Estate, Finance, and Local continue to have high growth and large numbers of visitors




US Category Growth by Page Views Sept 2005-Sept 2006
Page Views (billions)
60
50
40
30
20
10
0
□ Sep-05
▨ Mar-06
■ Sep-06
Community
Retail
Entertainment
Bus/Finance
Auctions
Sports
Regional/Local
Travel
Music
Classifieds
Real Estate
Health
Career Svc & Dev
CompShopping

World wide:

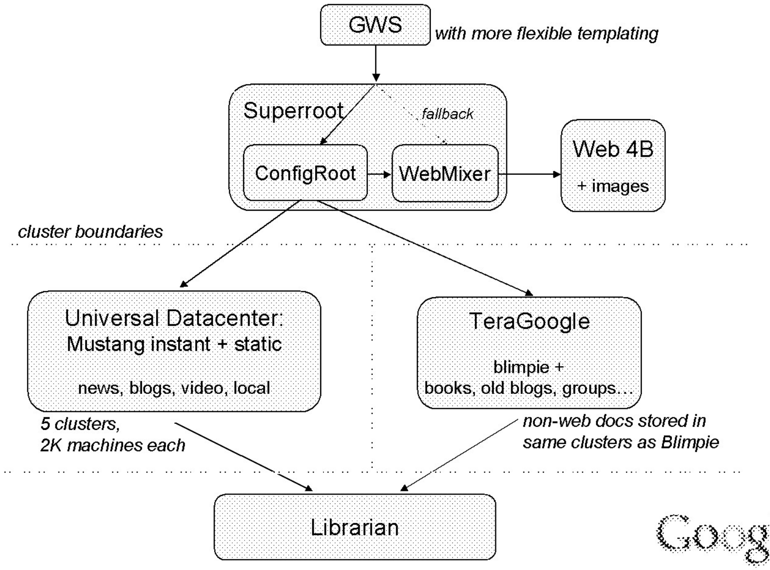

US Category Growth by Page Views Sept 2005-Sept 2006
Page Views (billions)
60
50
40
30
20
10
0
□ Sep-05
▨ Mar-06
■ Sep-06
Community
Retail
Entertainment
Bus/Finance
Auctions
Sports
Regional/Local
Travel
Music
Classifieds
Real Estate
Health
Career Svc & Dev
CompShopping



Appendix: Universal Architecture

We will use Librarian for index updates and all teams will be required to add live data updates to Librarian. We'll maximize the use of Teragoogle for data storage in order to minimize the use of machines. All other data will be stored in Mustang. The Mustang instant layer allows for frequently updating types of content and documents (like blogs and news). Superroot will support merging and re-ranking of data from different corpora. We will provide a standard mixing algorithm so that all types of content can be included. Finally, we must invest in a "Universal Front End" system that will enable the UI variety we anticipate. The Universal Front End will be a refactoring of GWS so that we are able to have incremental launches. It is important that this project get staffed and underway soon.

Universal serving diagram



8

International Market Share

- Russia: significant product improvements to core search and more aggressive localization of products has yielded improvements to search market share (9/2005: Yandex 47%, Google 15%; 9/2006: Yandex 46%, Google 22%; source: Spylog.ru)
- Japan: Google products & share improving. (1/3 of *all* Google mobile page views now come from Japan, followed by US (1/6), Google Maps Japan PVs has doubled from 8M/day to 16M/day in the past 3 months. Trails only US. Search share is flat.) Partner network under serious threat by Yahoo!Japan and impending launch of MSN AdCenter. (Lost Excite and Biglobe was a difficult renewal) MSN will likely offer large guarantees.
- China: market share slipping further; 2005 (Baidu: 48%, Google 33%), 2006 (Baidu: 62%, Google 25%); source: CNNIC, Annual China Search Engine Report, 2005/2006
- Korea: Naver (65%) and Daum (19%) dominate the search market with Google far behind (<2%); source: Korean Click