R&D at Baidu

Baidu is the leading Chinese language search engine on the Internet with a 44.7 percent market share. (Red Herring) They are regularly listed by Alexa as one of the most popular sites on the Internet. Their recent and highly successful IPO on the NASDAQ stock exchange has catapulted them to international renown. (Standard & Poor's) Although primarily operating in China, including conducting all research here, Baidu as an Internet company and thus must compete globally. Its competitors can be divided into two categories. Firstly, other search engines such as Yahoo, Google, and MSN.com. Secondly many other Chinese Internet properties such as Sohu.com and Sina.com against which it competes for advertising dollars. Advertising including 'pay for placement' is the primary source of revenue for Baidu. In Baidu's most recent quarterly report, online marketing revenue "was $10.6 million, an increase of 188.5 percent from the corresponding quarter last year." (Red Herring)

On October 20th 2005, Danna Zhu and I were lucky enough to interview the Director of Engineering, Guo Dan and one of his Senior Engineering Managers, Jeff Tang. The interview was conducted in both English and Chinese. Guo Dan in particular was more comfortable in Chinese so the questions were translated by Danna into Chinese and the answers as necessary were translated back into English. Some of the impromptu discussion took place in English and there was abundant use of English technical terms.

Guo Dan is one of the original founders of Baidu. At Baidu research is divided into two major areas. One area focuses exclusively on Baidu's spidering activity and web indexing efforts. This is referred to as the Search Group. Guo Dan heads up the other side, which they call Search Channels, which basically is all the other features of

Baidu.com, including the searching and indexing of MP3s, discussion forums, and news sites.

Jeff is a Senior Engineering Manager at Baidu. He was recruited to move back to China from the Silicon Valley. In the Valley he had worked for Netscape, AOL, and several other smaller startup companies. He is primarily focused on ERP and workflow improvements both inside of R&D and on how R&D integrates into the greater Baidu operation. Having lived abroad Jeff's English was particularly strong which was a significant factor in Baidu making him available to be interviewed.

Danna and I interviewed first Guo Dan and then Guo Dan and Jeff together, Danna was invaluable due to her native Mandarin speaking abilities. It was decided Danna would focus on the difficulties Baidu has in attracting and retaining high quality R&D staff while competing directly with deep-pocket American firms such as Google, Microsoft, and Yahoo. In addition, that I would focus on how R&D at Baidu relates to their business strategy particularly any international expansion plans.

Guo Dan described Baidu as a technology company first and foremost and thus R&D is at the heart of their operation and always will be. Five out of eight original Baidu employees were trained as engineers. R&D investment by the company is 12% of yearly revenues. In their most recent quarter "The company's research and development expenses amounted to $1.4 million." (Red Herring) The culture of Baidu and its non-traditional workplace and relaxed atmosphere is crucial to attracting and retaining high quality employees. One of the quirks of Baidu is you are allowed to bring your pet to work as long as it isn't a cat. The CEO is allergic to cats.

One of the key questions I put to Guo Dan was whether technology was developed and pushed onto the market or whether market demand or a niche was identified then R&D was directed to develop a product. At Baidu it seems to be a mixture of both with approximately 50% of products developed as pure R&D and the other half developed to fill a particular niche or request from advertisers or end-users.

I also inquired whether improving the quality of the data indexed by Baidu was the focus or whether increasing the amount of data indexed by Baidu was the primary driver behind R&D for the Search Group. Guo Dan was adamant that improving the quality of the index is the focus of R&D. Quality includes ensuring the freshness of the results, eliminating duplicate data, and combating the spammers who attempt to influence the search results for their own financial gain. (Search Engine Spam?) Guo Dan revealed that Chinese spammers are much more clever and dedicated than American spammers and thus spamming is possibly a far bigger problem for Baidu than it is for say Google. Baidu has the capability to update their index very rapidly in a few days if necessary.

Every feature developed by Baidu is aimed at advertisers. Everything is free to the end-user. The goal then of R&D is to develop useful tools and services for Chinese consumers and recoup development costs and generate future revenue streams through the sales of ads. This includes accepting payment for placement in the search results a practice largely pioneered by Inktomi (Spring), now part of Yahoo.

Although there are currently no official plans to expand Baidu's services to languages other than Chinese, there is little need to. China will soon become the country with the most citizens online. (Einhorn) This does not include the significant number of Chinese-speaking web surfers living in places like Taiwan and Singapore or the million plus Canadians of Chinese ancestry. (Overseas Chinese) The sponsored search market, which is Baidu's primary revenue source, is growing quickly in China. "Piper Jaffray estimates that it will grow to $1 billion by 2010. It is currently estimated to be $134 million." (Red Herring) That said Baidu's spider works automatically and needs to be able to consider and handle languages other than Chinese. The Baidu spider has visited my own humble website (http://www.muschamp.ca/) which is almost exclusively written in English and Baidu.com itself can handle queries in English. So although China and the Chinese language remain the focus, as Baidu indexes more and more of the web it will continue to expand the portion of its index that is in languages other than Chinese.

Other reasons Baidu continues to focus on the Chinese market is their ability to compete really well for Chinese speaking staff, the scarcity of individuals who truly understand search engines, particularly as they relate to parsing of natural Chinese, and the significant training costs to hire non-Chinese speaking R&D staff and integrate them into their existing operations.

The uniqueness of Chinese, which is compounded online, is a hurdle Baidu has had to overcome. The problems presented by the two rival character sets, the numerous competing encoding standards, and the inherent difficulty parsing natural Chinese particularly things such as idioms and proverbs now provides a barrier to entry to competitors such as Google and Microsoft.

An illustrative example is provided by the phrase 不三不四 bù sān bù sì. Even to someone who's Chinese is extremely limited I recognize these three characters. The phrase literally translates into "not three not four". However according to the dictionary when taken together they mean "dubious" or "shady". Furthermore several Chinese people have told me that this phrase is considered rude. It is this extra context provide by the Chinese characters particularly when combined together that makes indexing and even more so analyzing Chinese text extremely difficult to do algorithmically.

Furthermore, there is also no requirement to include spaces between words in Chinese or Japanese text though the practice is becoming more common. As a result sentences can be written “Iwentouttogetacupofcoffee.” Sentences such as this are extremely tough for non-native speakers to read and for computers to parse. The common practice of scanning forward and looking for ‘spaces’ and then assuming a group of characters represents a word is not always possible. “Natural language” as Guo Dan termed it, is thus more than the some of the characters and the spaces between them. This fact also explains why India, an offshore IT powerhouse, cannot be tasked with developing software for the Chinese market. Although capable programmers the fundamental differences presented by Chinese particularly in written form make it extremely difficult for non-native speakers to manipulate algorithmically.

Chinese people also have fundamentally different preferences on how they use the Internet and their expectations of it. This can be observed comparing Chinese portals such as Sohu.com to American counterparts such as Yahoo.com. The design of the Chinese site is considerably busier with all manner of blinking, flashing, and moving parts. A Western person finds these extra visual featurs annoying while Asians seem to have a higher tolerance for it. As the busiest aspects are often advertising, you can understand why a company like Sohu is reluctant to go with a more elegantly designed page.

Asian web surfers also seem to show a greater tolerance for long webpages. When online advertising was first introduced, advertisers wanted to be “above the fold” a term borrowed from print ads. It was thought users would be unlikely to scroll down so as much advertising and content was crammed into the top of the page as possible. Another factor that contributed to this practice, was browsers of the time often rendered the top of the page first.

Baidu.com is visibly similar to Google.com a website often praised for its simple interface. Everything is above the fold and advertising is kept to an absolute minimum. However when you click on a search result in Baidu it spawns a new window which is a preference of Chinese web surfers, where as Western web surfers prefer the opposite. This preference is so pronounced, queries to Google.com from an IP address inside China will spawn a new window when you click on a result. Yahoo.com has not chosen to alter its site to suit this Chinese consumer preference yet.

Baidu and Google dominate the Chinese internet search market. A recent report by the China Internet Network Information Center reveal substantial difference between who is using the two rival companies. According to an English summary of the report on News.com:

> “Google is most frequently visited for enterprise products, business opportunities, transportation services and travel searches.
>
> Study author Lu Weigang said Google, in contrast to Baidu, tends to attract high-end users--those who are well educated and have relatively high incomes. Baidu is favored by students, who account for a relatively

> large part of China's search population, according to the report. About 40 to 50 percent of Baidu users are students, the report said." (ZDNet China Staff)

Further complicating the relationship between the two rivals is the fact Google holds a 2.8% minority stake in Baidu. It is rumored that both Yahoo and Google made acquisition overtures towards Baidu. So far Baidu has chosen to remain independent. The reasoning behind this is to better position themselves as a 'national champion' to curie favours from the Government of China.

I also got the impression while talking to Guo Dan, that Chinese web searches tended to use more natural language queries as opposed to simply entering a keyword or short phrase as done in the West. This further emphasizes the importance to Baidu of algorithmically being able to understand and index natural Chinese.

Baidu has also invested heavily in technology to extract text from non-textual sources such as MP3s and videos. Additional focuses of R&D along these lines is on information filtering. Baidu currently has the ability to index documents in formats other than .html, this includes .doc and .pdf formats for instance. How Baidu does this surprised me when I asked about it, first the document is converted to plain text, then the plain text is indexed using Baidu's primary search technology. Baidu also has the ability to adjust its algorithm to place greater emphasis on the timeliness of data, this ability is currently used in their News search. I asked if Baidu had any plans to provide a search engine specifically for searching weblogs or blogs. This is something they are considering and they do not anticipate it being particularly challenging task. Baidu stated they could just use their existing algorithm and alter the emphasis place on timeliness, just like they do for their current news search and then focus their indexing and spidering activities exclusively on the blogosphere.

Some of the answers I received from Baidu surprised me. I can understand how they want to concentrate all R&D activities in China and their need for native Chinese speaking engineers to work on their Chinese language search algorithms. I'm also familiar with the concept of a 'national champion' and how China in particular is building towards these. The importance of this point was reiterated by Jeromy Xue 薛军 of the Tsinghua Science Park Venture Fund in a recent presentation to our MBA class. However their ignoring of other CJK (Chinese Japanese Korean) countries where their expertise in ideogrammatic search could be leveraged may prove shortsighted. Their primary American rivals are in all these markets and with the exception of Korea are first or second. (Korea Internet E-Commerce Toolbox FAQ)

Although flush with cash from their successful IPO, there seems to be some dissension both within and outside Baidu on where they should focus their efforts. It is clear from talking to Guo Dan and Jeff that China will continue to be the focus of the vast majority of Baidu's research and marketing efforts. The heavy reliance on P4P advertising, Pay for Performance as Baidu likes to term it, for revenue is leading to search results where the top places are all basically advertising. This may be a contributing factor in the more educated Chinese preferring Google. Baidu's relationship with Google, who they both admire and copy, even without publicly admitting it, but also see as their greatest rival bears watching. Although the Chinese market is large, even if they continue to be the leading search engine within it, they will be no more than a regional player compared to the Yahoo's, MSN's and Google.

References

Associated Press, . "IPO Ready: Chinese Google Wannabe." Wired 31 Jul 2005. 05 Nov 2005 <http://www.wired.com/news/business/0,1367,68376,00.html?tw=wn_story_related>.

Einhorn, Bruce . "China.Net." Business Week Online 15 2004. 04 Nov 2005 <http://www.businessweek.com/magazine/content/04_11/b3874012.htm>.

"Korea Internet E-Commerce Toolbox FAQ." Export.Gov US Government Export Portal. US Commercial Service. 05 Nov. 2005 <http://www.export.gov/sellingonline/Internet_Environment/Korea.asp>.

McKay, Muskie. "Baidu: China’s Search Engine." Muskblog. 31 Oct 2005. 04 Nov. 2005 <http://blog.muschamp.ca/>.

"Overseas Chinese." NationMaster. 04 Nov. 2005 <http://www.nationmaster.com/encyclopedia/Overseas-Chinese>.

Red Herring, . "Baidu Continues to Please." Red Herring 26 Oct 2005. 05 Nov 2005 <http://www.redherring.com/Article.aspx?a=14193&hed=Baidu+Continues+to+Please&sector=Capital&subsector=PublicMarkets>.

Standard & Poor's, . "Baidu.com Soars in IPO." Business Week 5 Aug 2005. 05 Nov 2005 <http://yahoo.businessweek.com/investor/content/aug2005/pi2005085_4630_pi004.htm>.

"Search Engine Spam?." O'Reilly Radar. 23 2005. O'Reilly Media, Inc.. 05 Nov. 2005 <http://radar.oreilly.com/archives/2005/08/search_engine_s_2.html>.

Spring, Tom. "Are Search Results for Sale?." PCWorld.com 28 2002. 04 Nov 2005 <http://www.pcworld.com/news/article/0,aid,86884,00.asp>.

ZDNet China Staff, . "Baidu, Google dominate Net search in China." News.com 30 Aug 2005. 05 Nov 2005 <http://news.com.com/Baidu,+Google+dominate+Net+search+in+China/2100-1038_3-5844468.html>.