

Vivísimo White Paper: Corporate Intranets

Estimating the Cost Savings with Vivísimo Document Clustering on Corporate Intranet Sites

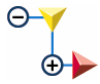
©2003

We provide a method to estimate the savings in employee time that could be expected by clustering the search results at a corporate intranet site.

The method builds on reports that the average user gives up on searching after about 12 minutes if a solution is not found. This is used to estimate that clustered results allow users to examine nearly double the number of relevant documents than in the case of result lists. Also, a clustering approach brings into potential view those documents that would be buried deep within a result list.

Combining these observations leads to estimated savings of more than \$1 million per year, under reasonable values for user behavior and employee salaries, and for a department site that sees a mere 2000 searches per day. There are additional benefits that are much more significant for the corporation, but which are not easily quantified, so we do not include them in our modeling.





Overview of Vivísimo Document Clustering

Vivísimo's clustering products help enterprises organize information from anywhere, any time, in any language without the endless cost and complexity of building information taxonomies. Vivísimo's technology is on display for web searching and for searching corporate, government, news, etc. sources at its corporate site <http://vivisimo.com>.

The Vivísimo Clustering Engine™ does not crawl or index a document collection. It organizes the outputs of other search engines: URLs, titles, summaries, and meta-data if available and desired.

The core Clustering Engine technology is called document clustering, which is the automatic organization of documents into spontaneous meaningful groups. Document clustering methods never need to touch or know about the larger collection from which search results are taken, or undergo any other pre-processing steps. Organizing the search results occurs *just before* a user is shown the long list of search results.

The final output is a hierarchy (or tree) on the left of a split screen with the search results on the right. The interaction is based on the familiar and good Microsoft Explorer style of interacting with a file system. A search result *is not* constrained to fit within a single tree location, since individual search results could reflect multiple themes.

For users, seeing clustered search results has three benefits:

1. Brings into easy view - by following the labels – those search results that otherwise would remain invisible because they are far down the list.
2. Leads to effortless knowledge discovery: a user learns the types or subtopics of available information relating to the query.
3. Provides context: places related documents within a single (sub)folder for joint viewing.

All of these have significant implications for a user's search productivity.



Effects of Search Frustration

A study reported by ZDNet.com found that people will not search for long on the web. There is a limited average time they will spend before giving up, or becoming very upset with the search technology available to them.

“On average, Web-rage is uncaged (sic) after twelve minutes of fruitless searching, although about seven percent of the 566 people surveyed by Roper Starch Worldwide say ire starts rising within three minutes.”
ref: <http://www.zdnet.com/zdnn/stories/news/0,4586,2667216,00.html>

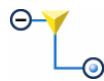
When a user becomes frustrated searching the *internet*, he or she has three usual options: try another search engine, re-formulate the search, or quit. There is little ill will toward the search engine since users often realize that the internet *content* is partly responsible.

However, when users search a corporate intranet site, they rightly believe that the corporation controls both the content and the search engine technology. If they cannot find what they seek, their confidence in the corporation suffers. We contend that:

1. A user becomes more impatient as the search drags on, increasing the frustration already felt from the problem that occasioned the search in the first place.

2. A reasonable estimate of an average search time is roughly 12 minutes, based on the "12 minutes to search rage analysis" and the recourse of telephoning customer support directly. It won't be much more than 12, and our model would predict that smaller numbers (e.g., 9 minutes) would lead to even greater savings.
3. Poor experiences will lead users to bypass the intranet site in the future and to opt directly for other ways to obtain information, defeating the intranet's purpose.

Some of these negative consequences are significant but hard to quantify, especially in dollar terms. The next section makes use only of the 12 minute average search time phenomenon, which will lead to arguably large dollar savings credited to clustering. However, the largest savings are likely to accrue from those factors that we are unable to quantify.



Clustered Search Results Lead to More Document Views

Based on the search rage phenomenon, we can estimate how many search results can be "pulled" (clicked on, examined) by the average user using either the traditional ranked list representation or Vivísimo document clustering. Let us define the following parameters with some typical values for them:

R = 25 : Number of hits returned on one search page when using ranked lists

Cavg = 12 : Number of top-level folders returned by Vivísimo

T-Analysis = 5 sec. : Average time to judge the relevance of a single document's summary in a ranked list, or the relevance of a single folder group found by Vivísimo

T-Docs = 60 sec. : Average time to read a document obtained by clicking on a document summary, and to decide whether the document solves the user's problem

NUMBER OF DOCUMENTS VIEWED IN 12 MINUTES					
Ranked-List output (without Vivísimo)					
Description	Formula	Search	Docs Examined	Time (sec)	Running Total (min)
Create search string/ Returns number "R" hits		1		20	0.33
Analyze results based on document summaries	(R x T-Analysis)			125	2.42
User pulls on average 2 of the returned docs	(2 x T-Docs)		2	120	4.42
Consult next series of search results		1		5	4.50
Analyze results based on document summaries	(R x T-Analysis)			125	6.58
User pulls on average 2 of the returned docs	(2 x T-Docs)		2	120	8.58
Consult next series of search results		1		5	8.67
Analyze results based on document summaries	(R x T-Analysis)			125	10.75
User pulls on average 2 of the returned docs	(2 x T-Docs)		2	120	12.75
Totals		3	6		12.75
With Vivísimo's Clustered Output					
Description	Formula	Search	Docs Examined	Time (sec)	Running Total (min)
Create search string/ Returns number "R" hits		1		20	0.33
Analyze Folders, Decide which is of interest	(Cavg x T-Analysis)			60	1.33
Analyze Results in the selected Folder	(# of Docs x T-Docs)		11.4	684	12.73
Totals		1	11.4		12.73

(Above we have assumed that clustering operates on 200 results returned by the search engine, which is a typical number, but undesirably small if the results will be clustered.)

The basic intuition is that a user will look at an average 12 top-level folder groups returned by Vivísimo clustering, pick the most relevant one, and click on and examine about 11 documents there before the time-to-frustration is exhausted. In the ranked-list case, only 6 documents can be clicked on and examined. The difference between the two cases is because the ranked list makes the user waste time skimming and discarding many irrelevant documents on successive pages of the search results.

Our further analysis will only make use of this difference in the number of documents that an average user can examine closely.

As stated before, there are benefits from clustering that do not easily lead to quantifiable results, so we can only mention but not use them in our subsequent analysis. Clustering brings into potential view documents that are far down in the results list (ranked 137, say, so practically invisible), but could be what the user actually needs. Typical search engines return at most 200 results, but Vivísimo clustering can easily deal with 500 results and thus brings many more documents into potential view.



[Comparative Chances of Solving a Problem Requiring Intranet Search](#)

It would be ideal to either calculate or empirically measure the comparative chances of solving a support problem by ranked lists or by clustering. However, too many unknown parameters are needed for calculation, and experiments with actual users are outside the scope of this white paper. Instead, we follow a relatively simple approach: assume that the overall chance of solving the problem is 50% with the ranked approach. If we assume that each document view has an independent chance of solving the problem, this implies (see the appendix) a 72% chance to solve the problem with clustering.

It may be objected that if a ranked list is very well ranked based on relevance criteria, then ranked lists will do better than this ratio predicts. However, in practice it is difficult for a ranking algorithm to “know” the user’s intent based on the sparse search queries that are seen in practice. Moreover, longer queries can even be counterproductive. For example, if the needed solution document is generic to many product models or product lines, a search engine may rank this document poorly if the search query specified a particular product model.

It should be added that the authority-based ranking methods responsible for successful search engines such as Google do not fit the circumstances of customer support sites, which lack the rich patterns of cross-hyperlinking on the web that underlie the calculations carried out by those methods.



[Cost of Personnel](#)

We use an average employee cost of \$60 per hour, including non-salary expenses.



[Estimate of Savings in Support Cost with Document Clustering](#)

If there are no answers in an intranet site, then clearly clustering cannot help. We are unable to estimate what percentage of user problems have stored answers, so our analysis is based on the assumption that every problem has an answer, or, that the number of daily searches we use corresponds to those searches which actually possess an answer.



Vivísimo

We have the needed parameters to proceed with a simple model.

- Empl = number of employees in unit
- Info = number of information needs per employee/day that lead to trying to find answers on corporate intranet
- $P_{\text{clustered}}$ = percentage (fraction) of users who find an answer using clustering
- P_{ranked} = percentage (fraction) of users who find an answer using ranked lists
- Time = average time (minutes) to solve problem using other sources
- Cost = employee cost per minute

The savings per day is directly proportional to the number of users who solve their problem with clustering but not with rankings. Earlier we assumed that P_{ranked} is 0.5 (50% solution rate) implying (see appendix) that $P_{\text{clustered}}$ is 0.72 (72% solution rate). Thus the formula for annual savings is:

$$\text{Savings/yr} = \text{Empl} \times \text{Info} \times (P_{\text{clustered}} - P_{\text{ranked}}) \times \text{Time} \times \text{Cost} \times 365$$

Corporate Cost Savings: Intranet Installation	
Empl (Number of employees using Intranet Search)	1,000
Info (# of daily searches per user)	2
<i>Total Searches per Day</i>	
	<i>2,000</i>
Successful Searches per day (ranked list)	1,000
Successful Searches per day (Document Clustering)	1,440
<i>Additional Successful Searches w/Vivísimo</i>	
	<i>440</i>
Time (Minutes Saved per successful search)	10
Total \$ amount saved per successful search (\$60/hr)	\$10
Daily Savings per Intranet Customer	\$4,400
Number of 'work days' per year	230
Annual Savings per Intranet Customer	
	\$1,012,000

With these quantities, the annual savings is more than a million dollars. *The annual savings per employee is \$1,012.*

Readers may use our formula to explore various scenarios. For example, 4,000 searches per day (instead of $\text{Empl} \times \text{Info} = 2000$) would double the estimated savings. A 50% reduction in the cost of alternative information sources would halve the savings, which would remain large.



Conclusion

This white paper has provided a simple plausible model that estimates the expected monetary savings from introducing Vivísimo document clustering on a corporate intranet site. The annual savings are around \$1,012 *per employee* using reasonable values for parameters that describe user behaviors. A typical large department would see upwards of a million dollars in annual savings.

We can point out additional benefits of clustering which are more difficult to quantify but are of clear value. Actually, these unquantified gains are *much more significant* than the time savings in our model.

- An enhanced user experience with gains for the image and reputation of the intranet site.

- A decrease in the fraction of users who lose faith in the utility of an intranet site and instead directly turn to other information sources.
- Our model doesn't include the *doubled time savings* when a user consumes the time of other employees in order to solve his information need.
- The ability for a user to learn about company operations by viewing the taxonomy (or hierarchy) that is returned as the output of every search. For example, doing a search on a specific corporate customer may turn up a taxonomy of the interactions with that customer that are recorded in the intranet's document base. This easy mechanism of knowledge diffusion is probably the biggest benefit.
- A reduced need on the part of the intranet site engineers to improve the ranking of search results, which is a harder challenge than on the web.

Finally, it should be noted that *just in time clustering* does not interfere with separate advances in intranet sites, e.g., increasing the number of search engines or knowledge bases, filtering outputs according to the available user profile, improving the summaries that are associated with each document, etc. Regardless of these advances, the number of information items relevant to user requests will keep exceeding the number that users will be willing to examine. Hence document clustering can always be used with advantage.



Appendix

Let's say that the probability of success with ranked lists (P_{ranked}) is 50%, that 6 documents ($\#Rdocs$) can be viewed in the ranked list, and that 11 documents can be viewed with a clustered output ($\#Cdocs$). Assume that each document view has an independent probability p of being the solution. Then

$$P_{ranked} = 0.5 = \sum_{i=1}^{\#Rdocs} p(1-p)^{i-1}$$

With a little algebra, the expression can be re-written as:

$$P_{ranked} = 0.5 = 1 - (1-p)^{\#Rdocs}$$

Plugging in the numbers and solving for p yields a value of 0.11. Then plugging $p=0.11$ and $\#Cdocs = 11$ (clustered output) into the formula

$$P_{clustered} = 1 - (1-p)^{\#Cdocs}$$

gives a value of 0.72.

Thus, if P_{ranked} is 50%, then $P_{cluster}$ is 72%. That is, the chance of solving the problem with $P_{cluster}$ is around 44% higher than with P_{ranked} .

whitepapers@vivisimo.com

Copyright © 2003 Vivísimo, Inc.