

Inktomi® Corp. : Web Search Relevance Test

Test report prepared under contract from Inktomi Corp.

Executive summary

Inktomi Corp. commissioned VeriTest, a division of Lionbridge Technologies, Inc. to conduct a Web Search Relevance Test comparing the Inktomi search engine to the following search engines:

- Google
- WiseNut
- FAST
- Teoma
- AltaVista

The goal of the testing was to compare the relative quality of the top 10 Web documents returned by each of the search engines under test as a result of entering specific search criteria. The URL returned as a result of a specific query was then judged according to whether or not it was relevant based on the judging guidelines. We then applied several different weighting formulas to the raw data to determine which of the search engines tested provided the most relevant information.

To generate the queries used during the testing, Inktomi provided an actual Web query log of more than 7 million queries. Inktomi cleaned the log to remove links related to adult material, queries for URL's, nonsense queries, queries containing non-English characters and duplicates. We monitored and audited the process to insure that the process did not bias the results. We randomly selected 100 queries from the resulting set of 1 million cleaned queries. We ran each of the 100 queries on each of the 6 test engines and recorded the links or URL's to each of the top 10 web documents returned by the search. Sponsored links were ignored.

We submitted the links or URL's returned as a result of each query to a panel of three judges. We hid the source search engine and the rank of each document returned by the search from the judges. The judges were given a set of guidelines and asked to score each of the web documents to which the URL's or links pointed as acceptable or not acceptable based on the relevance of the document to the query. See the Methodology section of this document for more details on the judging standards.

Key findings

- ❑ We found that when we examined raw un-weighted relevance scores, Inktomi at 54.33%, scored slightly ahead of Google at 53.23%. WiseNut, Teoma, AltaVista, and FAST followed, grouped closely together with scores of 42.57%, 42.50%, 40.73% and 39.10% respectively.
- ❑ When we applied the $1/r$ coefficient of weight, we found that again the scores that we awarded to Inktomi and Google were very close, with Inktomi slightly higher at 509.0 points. Google scored 502.5 points and WiseNut, Teoma, AltaVista, and FAST received totals of 429.4 points, 428.1 points, 405.7 points and 411.6 points respectively.
- ❑ When we applied the $1/\sqrt{r}$ coefficient of weight, we found that again the scores that we awarded to Inktomi and Google were very close, with Inktomi slightly higher at 842.0 points. Google received a total of 831.0 points. WiseNut received a total of 681.8 points. Teoma received a total of 682.6 points. AltaVista received a total of 651.3 points. FAST received a total of 642.5 points.
- ❑ We found that when we examined the percent of web documents judged acceptable, position by position, we found that all six search engines scored well in position number one but the rate of acceptable judgments in positions two through five fell more rapidly for WiseNut, Teoma, AltaVista and FAST than for Inktomi and Google.

We then tallied a raw score and two different weighted scores based on position in the return list, to compare the relative quality of the results returned by each of the search engines. We found that Inktomi and Google consistently scored higher than the other four engines in both the raw scores and the weighted scores, generally with Inktomi scoring slightly higher than Google.

We found that the scores of the other four search engines, WiseNut, Teoma, AltaVista, and FAST were also closely grouped but somewhat lower than those of Inktomi and Google. We found that when we examined the scores on a position by position basis, that the number of web documents that appeared in the number one position on the list of links returned by the query that our judges found acceptable were similar for all six engines with Inktomi scoring slightly higher than WiseNut. However we found that the rate of acceptable documents fell more rapidly in positions two through ten for WiseNut, Teoma, AltaVista and FAST than it did for Inktomi and Google. See the Test Results section of this document for a more detailed examination of the scores.

Testing methodology

Inktomi Corp. commissioned VeriTest, a division of Lionbridge Technologies, Inc. to conduct a Web Search Relevance Test comparing the Inktomi search engine to the following search engines:

- Google
- WiseNut
- FAST
- Teoma
- AltaVista

The goal of the testing was to compare the relative quality of the top 10 Web documents returned by each of the search engines under test as a result of entering specific search criteria. The URL returned as a result of a specific query was then judged according to whether or not it was relevant based on the specific judging criteria. We then applied several different weighting formulas to the raw data to determine which of the search engines tested provided the most relevant information.

The following sections provide the details of the methodology used to conduct these tests

Query Selection

In order to test the relevance of the search engines, we needed a set of queries to send through each search engine under test. Inktomi provided us with an actual search engine query log containing over 7 million queries to use during testing. Inktomi used previously developed software to flag and remove searches for adult material, URL's, IDP Meta queries, queries that contained non-English characters, duplicates and non-sense queries. Inktomi provided us with the original log with flags for each query eliminated to indicate why their software had eliminated the query. We independently audited the resulting subset of one million queries by selecting random subsets of the original group and repeating the cleaning process by hand. We then compared our results with the Inktomi results to insure that the cleaning process had not biased the query pool. We selected from the 1 million a subset of 130 queries by first using a random number generator to assign a random number to each query. We selected the first 130 queries in the randomized list. We eliminated any that we felt were intended to locate adult material that had been missed by the original screening. We also eliminated any that did not produce at least 10 hits by each of the search engines in the test. We did not attempt to spell check or make any other judgments about the selected queries.

After selecting the final 100 queries to be used for the test, executed each of the 100 queries on each of the six search engines in the test. We recorded each of the top ten URL's returned by each engine. We ignored sponsored links.

A list of the queries and the URL's produced by each search were presented to a jury of three judges. The rank of the URL and the search engine that produced the URL were hidden from the judges.

Judging

After recording the top 10 URL's returned by each search engine, we setup a panel of three judges to determine the relevance of the returned content as it related to the original query from which it was generated. We gave each judge a list of the queries used to generate the URL's and a list of links to the Web documents that appeared at least once in the top ten returns from each of the searches. The search engine or engines that listed the URL and the rank of the URL in the listing were hidden from the judges. The judges were not allowed to consult on their decisions or to discuss any of the queries during the judging process.

We instructed the judges to look at each query and decide what the query meant or the specific information for which the query maker was looking based on a set of judgment criteria described in the following section of this report. Each judge then used a browser to view the Web document associated with each of the URL's returned as a result of a specific query. Each judge entered a judgment based on the content of the document and the standards for judging that we gave each judge prior to the start of judging.

The judges chose from three possible judgment values.

- Accept - For any document that the judge believes is a useful result for the query.
- Reject - For any document that the judge believes is not a useful result for the query.
- No judgment - For any document that cannot be evaluated (usually because the URL cannot be accessed).

Judgment Criteria

Inktomi and VeriTest agreed on the following set of judging criteria. Each judge was given the criteria prior to the beginning of the judging with out additional comment or discussion:

Analyze the purpose of the query. Accept any document that you deem both relevant and useful, given the purpose of the query. Reject any document, regardless of relevance, if you decide it is not a useful document for the query.

Note that document utility is a context-dependent variable, since the attributes that define usefulness will vary from one query to another. For example, the criteria you use to judge whether a document is useful for the query "Unified Field Theory" will differ markedly from the criteria you use for the query "Cindy Crawford." It's up to each individual judge to decide what constitutes a useful document for the query being judged.

We gave each judge the following written set of questions to ask when making their judgment. We gave the list of questions without additional comment or discussion.

- *If I were interested in the subject of this query would I bookmark this URL?*
- *If a friend of mine was interested in the subject of this query, would I email them this URL?*
- *If I only had access to one web page on the subject of this query, would this URL be a good choice? (Note: This question allows for more than one URL to be "a good choice")*
- *If I were writing a FAQ list on the subject of this query, would I include this URL?*
- *If I were publishing an article on the subject of this query, would this URL be useful in my background research?*
- *If I were writing a paper on the subject of this query, would I include this URL in my references?*

We also gave each judge the following set of grade equivalents to use as a guide.

- *Excellent = Accept*
- *Good = Accept*
- *Fair = Reject*
- *Poor = Reject*
- *Spam = Reject*
- *Porn = Reject*
- *Intl (if the language of the document is not appropriate, given the language of the query) = Reject*
- *Intl (if the language of the document is appropriate to the query, but you can't read the language) = No Judgment*
- *Down = No Judgment*
- *404 = No Judgment*
- *F (access forbidden) = No Judgment*

The judges entered a score for each web document in a scoring database.

Scoring

Once the judging process was complete, we calculated a raw score and weighted score to measure the relevance of each search engine tested. The number of relevant or acceptable scores was tallied for each query and in each position in the list of returned URL's. The percentage of relevant or acceptable pages in each position was calculated as well as an overall percent of relevant pages for each search engine.

We also calculated weighted scores for each engine based on two different weighting formulas. Weighting the scores allowed us to add the significance of the relative position of each URL in the list to the final score. The URL in the first position carried the most weight and the URL in the 10th position carried the least. The first weighted score was calculated using the formula $1/r$ where r is the rank to calculate the weighting coefficient. We calculated a second weighted score for each engine using the formula $1/\sqrt{r}$, where r is the rank, for the coefficient of weight. See the Test results section for a detailed report of the scores.

Test results

This section provides the details of the search engine relevance testing. For these testing, we performed 100 queries on each search engine and collected the top 10 Web document links or URL's from each engine, for a total of 1000 web docs to be judged. Each of three judges awarded a score of "Accept", "Reject", or "No Judgment" for a possible raw score of 3000 "Accepts" for each search engine. Please refer to the Testing Methodology section of this report for details of how we conducted the testing.

After completing the judging phase of the testing, we first compiled the raw scores for each search engine in the test assigning equal weighting to all URL's returned regardless of their position in the list of returned URL's

We found the results for Inktomi and Google were very similar with Inktomi scoring slightly higher. WiseNut, Teoma, AltaVista and FAST scores were closely grouped, somewhat lower than Inktomi and Google. Figure 1 shows a summary of these raw scores expressed as percent. Inktomi received the highest total raw score of 1630 “Accepts” out of a possible 3000 from the judges or 54.33% judged acceptable. Google followed closely receiving a total of 1597 “Accepts” or 53.23% judged acceptable. WiseNut received a total of 1277 “Accepts” or 42.57% judged acceptable. Teoma received a total of 1275 “Accepts” or 42.50% judged acceptable. AltaVista received a total of 1222 “Accepts” or 40.73% judged acceptable. FAST received a total of 1173 “Accepts” or 39.10% judged acceptable.

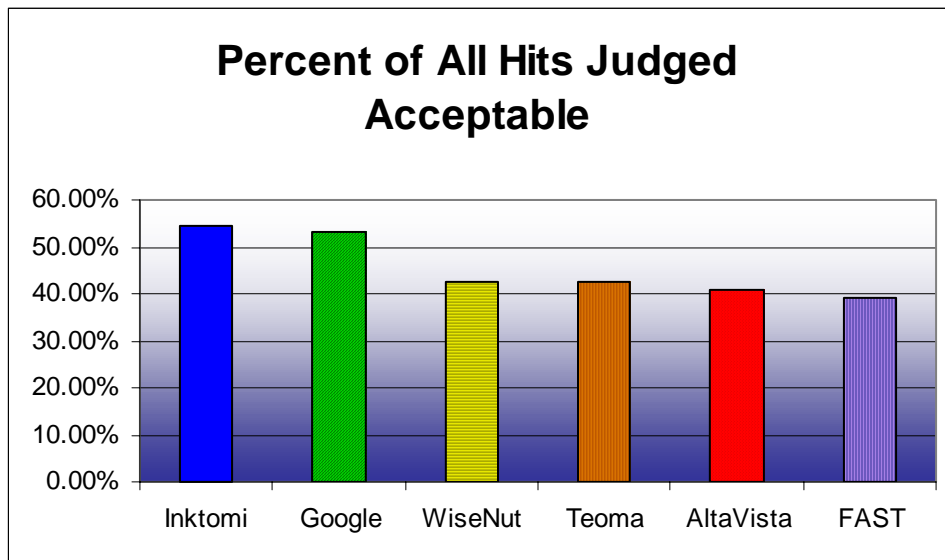


Figure 1: Raw Score expressed as a Percent of total possible score.

We also noted some interesting results when we examined the raw scores in a position-by-position basis. For example, when we looked at the percent of acceptable web documents in the number one position the scores for each engine were more closely grouped than the over all scores. Figure 2 shows the results for the first five positions. Inktomi received the highest score with 65.33% judged acceptable in the first position. WiseNut followed closely 63.67% judged acceptable. Google received a total of 61.67% judged acceptable. Teoma had 60.00% judged acceptable. AltaVista had a total of 56.33% judged acceptable in the first position. FAST received a total of 60.0% judged acceptable in the number one position. But as shown in Figure 2, Inktomi and Google continued to have acceptable rates that for the most part exceeded 50% in the next 4 positions. However the rate of web documents judged acceptable in the next 4 positions for WiseNut, Teoma, AltaVista and FAST on average fell below 50% with some slipping below 40% acceptable. Most

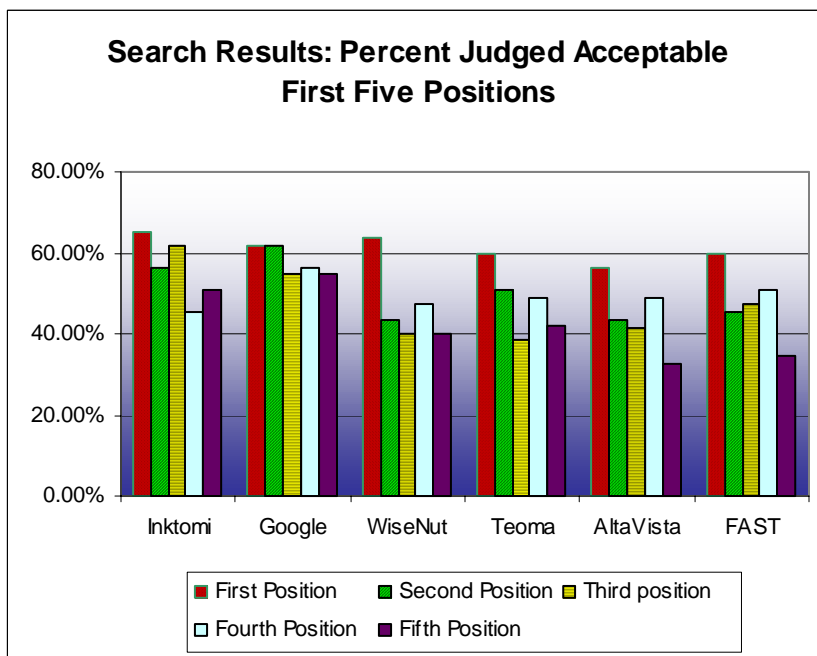


Figure 2: Raw scores of the first five positions on the search list.

notably, WiseNut had the second highest score in position one but scores fell significantly in positions 2 through 5. The chart in Figure 3 has a summary of these scores.

Position	1	2	3	4	5
Inktomi	65.33%	56.33%	61.67%	45.33%	51.00%
Google	61.67%	61.67%	54.67%	56.33%	54.67%
WiseNut	63.67%	43.67%	40.00%	47.33%	40.00%
Teoma	60.00%	50.67%	38.33%	49.00%	42.00%
AltaVista	56.33%	43.67%	41.67%	49.00%	32.67%
FAST	60.00%	45.33%	47.33%	51.00%	34.67%

Figure 3: Raw scores of the first five positions on the search list.

Next, we applied two weighting coefficients to the raw score to reflect the importance of rank in the score. First we awarded 1 point for each web document judged acceptable and then calculated the weighted score using the formula,

“weighted score = raw score*(1/r)” where r is the rank of the URL in the query hit list. For example, if a document was judged acceptable, and the document was in the first position in the query hit list, we awarded the search engine $1*(1/1)=1$ point. If the document in the second position in the query hit list was judged acceptable, we awarded the search engine $1*(1/2) = \frac{1}{2}$ point. We awarded $1/3$ point for documents in the third position on the list if the document was judged acceptable, etc. Again,

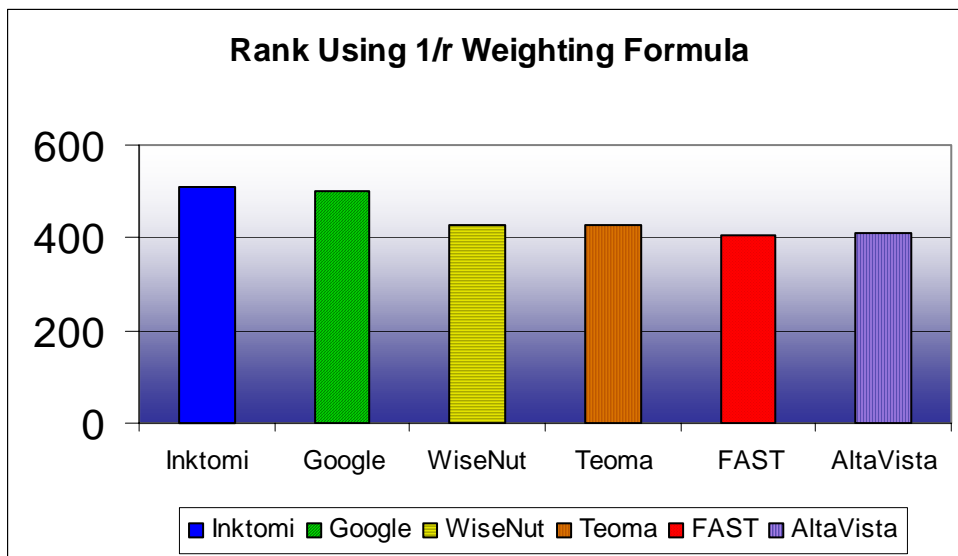


Figure 4: Weighted Score Using 1/r as the coefficient of weight.

Inktomi finished slightly ahead of Google with the other search engines grouped together with somewhat lower weighted scores. Figure 4 shows the ranking of each engine using 1/r as the weighting coefficient. Inktomi received the highest total weighted score of 509.0 points. Google followed closely receiving a total of 502.5 points. WiseNut received a total of 429.4 points. Teoma received a total of 428.1 points. AltaVista received a total of 405.7 points. FAST received a total of 411.6 points.

We applied a second coefficient of weight using the formula, “weighted score = raw score*(1/sqrt[r])” where r is the rank of the URL in the query hit list. For example, if a document was judged acceptable, and the document was in the first position in the query hit list, we awarded the search engine $1*(1/\sqrt{1})=1$ point. If the document in the second position in the query hit list was judged acceptable, we awarded the search engine $1*(1/\sqrt{2}) = .707$ of a point. We awarded .577 of a point for documents in the third position on the list if the document was judged acceptable, etc. Again, Inktomi finished slightly ahead of Google with the other search engines grouped together with somewhat lower weighted scores. Figure 5 shows the ranking of

each engine using $1/\sqrt{r}$ as the weighting coefficient. Inktomi received the highest total weighted score of 842.0 points. Google again followed closely receiving a total of 831.0 points. WiseNut received a total of 681.8 points. Teoma received a total of 682.6 points. AltaVista received a total of 651.3 points. FAST received a total of 642.5 points.

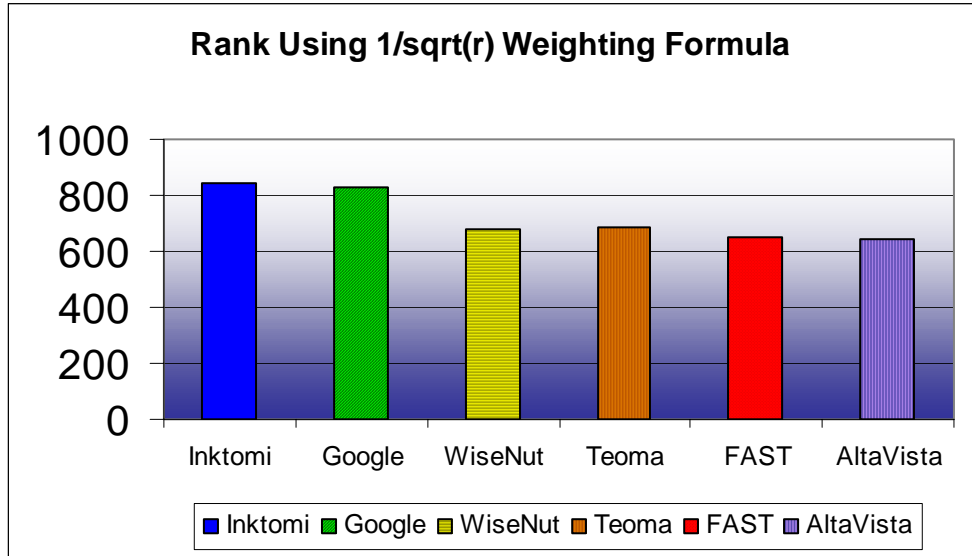


Figure 5: Weighted Score Using $1/\sqrt{r}$ as the coefficient of weight.

Appendix A

List of queries used.

2002 sparks bike rally
90.5 RADIO STATION
african american heritage
Airline Luggage
american sportfishing association2002
artex
baby names that start with j
bedbathandbeyond
bhangra album
big bend location
Bob Burnquist
boo williams
buckrail lodge
budweiser commercials
building and repairing computers
Cash Register software
clone cd
computer donation
elocon cream
equestrian gifts
Federal reserve Meeting
first And class And american And
credit And union
Gambino family web page
genitourinary
good morning am tv show
Greencastle IN economic development
greg iles
healthy diets for children
heart wise
Help wanted eastern long island
HIPAA Title II Subtitle F
history of the king cake
homes to buy in france
houston harvest
"How to create a jazz or a blues
phrase"
hydro power
in da club lyrics
Information on Lung Cancer
insert eps file to powerpoint
international study brussels
jason ligget
LANTANA RESTAURANT NJ
las mananitas
littman stethoscope
lotion dispenser for back
Louisiana Fairgrounds
Marriot senior Living
mary maxim
MARYLAND INSURANCE DEPARTMENT
match making india
mattress discounters
mccomb ms
mechanical engineering uk
Michael Bibby Fact
monteverde
msn public profiles
neverwinter online crack
New Found Glory
newkirk environmental
Newport High
NJ Superior Court
North Carolina Boys State
oboe clip art
Office Max Denver
Outlook express download
overseas job tax free
pendidikan
phone number reverse search
Photo of Esteli Nicaragua
planet names
Plumbing supplies rheem
POSTCODES
public auctions trenton ohio
rci home
relief stone accent tile
Republican National Com
RESTAURANT PAPER PLACEMATS
retail week
retriever data base
rhonda dotson
Robert Morris Associates
romantic hotels
secretary of state mississippi
south asia arts
spax
spiritual books for women
ST Lawrence collegein British
telus
TENNESSE AUTO DEALERS
the best karaoke machine
The east african standard
"United States defense contractor
fined"
university of maryland in europe
Ursa Minor
Vendor Evaluation Form
vsp
Weather Network
wild tymes restaurant
wrdu

yahoo maps
zx spectrum

VeriTest (www.veritest.com), the testing division of Lionbridge Technologies, Inc., provides outsourced testing solutions that maximize revenue and reduce costs for our clients. For companies who use high-tech products as well as those who produce them, smoothly functioning technology is essential to business success. VeriTest helps our clients identify and correct technology problems in their products and in their line of business applications by providing the widest range of testing services available.

VeriTest created the suite of industry-standard benchmark software that includes WebBench, NetBench, Winstone, and WinBench. We've distributed over 20 million copies of these tools, which are in use at every one of the 2001 Fortune 100 companies. Our Internet BenchMark service provides the definitive ratings for Internet Service Providers in the US, Canada, and the UK.

Under our former names of ZD Labs and eTesting Labs, and as part of VeriTest since July of 2002, we have delivered rigorous, objective, independent testing and analysis for over a decade. With the most knowledgeable staff in the business, testing facilities around the world, and almost 1,600 dedicated network PCs, VeriTest offers our clients the expertise and equipment necessary to meet all their testing needs.

For more information email us at info@veritest.com or call us at 919-380-2800.

Disclaimer of Warranties; Limitation of Liability:

VERITEST HAS MADE REASONABLE EFFORTS TO ENSURE THE ACCURACY AND VALIDITY OF ITS TESTING, HOWEVER, VERITEST SPECIFICALLY DISCLAIMS ANY WARRANTY, EXPRESSED OR IMPLIED, RELATING TO THE TEST RESULTS AND ANALYSIS, THEIR ACCURACY, COMPLETENESS OR QUALITY, INCLUDING ANY IMPLIED WARRANTY OF FITNESS FOR ANY PARTICULAR PURPOSE. ALL PERSONS OR ENTITIES RELYING ON THE RESULTS OF ANY TESTING DO SO AT THEIR OWN RISK, AND AGREE THAT VERITEST, ITS EMPLOYEES AND ITS SUBCONTRACTORS SHALL HAVE NO LIABILITY WHATSOEVER FROM ANY CLAIM OF LOSS OR DAMAGE ON ACCOUNT OF ANY ALLEGED ERROR OR DEFECT IN ANY TESTING PROCEDURE OR RESULT.

IN NO EVENT SHALL VERITEST BE LIABLE FOR INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH ITS TESTING, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. IN NO EVENT SHALL VERITEST'S LIABILITY, INCLUDING FOR DIRECT DAMAGES, EXCEED THE AMOUNTS PAID IN CONNECTION WITH VERITEST'S TESTING. CUSTOMER'S SOLE AND EXCLUSIVE REMEDIES ARE AS SET FORTH HEREIN.