



whitepaper

# microlanguage™ thematix™

the cost of conventional search technology compared with  
intelligent language processing

## abstract

The true cost of a knowledge management application is not the up-front costs of acquiring it, but the cost of paying knowledge workers to use it. Cost is thus a function of the quality of the information returned by the application. Knowledge management applications based on conventional search technology tend to return poor results because they can not differentiate between relevant and irrelevant information. Ultimately, the long-term low-cost solution is to replace conventional search technology with intelligent language processing technology that is able to distinguish relevant from irrelevant information: Microlanguage's Thematix technology.

## table of contents

the cost of lost productivity .....	3
the real cost of a solution: the whole truth and nothing but the truth.....	3
the cost of conventional search technology .....	5
thematix technology: the long-term low-cost solution.....	6

## table of figures

poor vs idealized search performance .....	4
--	---

## about microlanguage

Microlanguage develops and markets next-generation intelligent language processing technology and services to OEMs, systems integrators, and portal information providers. Our technology enhances the value of our partners' products and services to global 2000 technology, finance, healthcare, pharmaceutical, manufacturing, and national security domains. For more information on Microlanguage and Thematix, visit us on the web at [www.microlanguage.com](http://www.microlanguage.com), or call us at 610.995.1017.

## cost benefits of intelligent language processing: a \$2 billion opportunity

### the cost of lost productivity

According to a study on knowledge worker productivity, market research firm IDC<sup>1</sup> concluded that enterprise employees may be losing as much as three hours per day on fruitless information searches. In the context of the top 1,000 U.S. companies, each with an average of 1,000 knowledge workers, of the \$80 billion budgeted annually for knowledge worker activities, IDC estimates that \$2.6 billion is being spent on the search for indexed information. An average of 80% of that time, which translates to \$2 billion, is wasted searching through irrelevant information returned by conventional search technology.

This paper explains how can these costs be recovered.

### the real cost of a solution: the whole truth and nothing but the truth

If knowledge workers could get at the right information more quickly, then a significant portion of the lost productivity and the money spent on it could be returned to the bottom line. The goal is KNOWLEDGE MANAGEMENT, i.e. to keep track of all of a company's information assets and to access relevant information quickly and accurately. Many knowledge management applications provide a solution, but at what cost?

There are three costs associated with any solution: the initial cost of acquiring the product, the cost of finding information, and the cost of sorting through the information returned by the search. We look at each of these more closely to determine the real cost of a knowledge management solution.

**the cost of acquiring the solution.** The up-front costs can vary for solutions. However, the up-front costs do not tell us anything about the long-term costs of using the solution.

**the cost of finding information.** The cost of the downtime associated with waiting for a search application to return results was more important when computing power was limited. With the tremendous speed of current processor and memory technology, this question is largely moot: most any application can return information

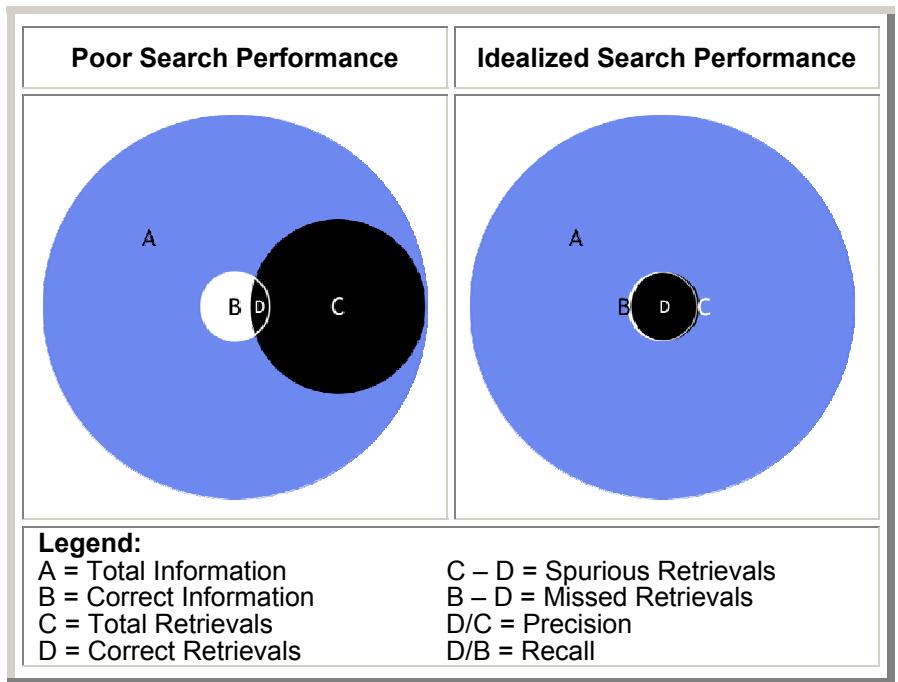
---

<sup>1</sup> Infoworld Magazine, January 7, 2002 issue.

quickly, and the differences between search times are trivial enough that no cost is associated with them.

**the cost of sorting information.** The fact is, conventional search technology returns lots of information quickly, but not necessarily the information that the knowledge worker was looking for. Unlike the one-time cost of acquiring a solution, this cost is on-going and measured in burdened man-hours. The following figure illustrates where costly knowledge workers are losing productivity.

## poor vs. idealized search performance



For each side of the figure, the outer circle A represents all of the information assets available. The inner circle B represents the relevant information that the knowledge worker is after. The circle C represents the information returned by a search. There are several important things to notice about C. For poor performance, the search misses much of the relevant information, and returns lots of irrelevant (i.e. SPURIOUS) information. With such results, the knowledge worker must spend time wading through all of the information returned in search of the relevant information. Moreover, a new search must be instigated in an attempt to gather the missed information, or worse yet, the knowledge worker has no idea how much relevant information was missed.

To clarify the relationship of these sectors to each other, we relate them to a common phrase: “the truth, the whole truth, and nothing but the truth.”

- Since circle B represents the total relevant information, we refer to B as THE TRUTH.
- THE WHOLE TRUTH is the totality of the desired information, i.e. all the relevant information that the knowledge worker has in mind when instigating a search. The standard term for this relationship between the desired information available and the desired information returned is RECALL. If the search returns all of the desired information, then the searcher has the whole truth.
- Furthermore, the knowledge worker wants the search to return only that information that is relevant and not undesired information that must be waded through to find the relevant information, i.e. NOTHING BUT THE TRUTH. The standard metric that expresses the relationship between desired information available and the undesired information returned is PRECISION. If the search returns only relevant information, then the searcher has nothing but the truth.

The more irrelevant information is returned, the more time the knowledge worker must spend sorting through it, and so the real cost of a knowledge management application is measured by its precision.

Ideal performance is when the retrieved information is both “the whole truth” and “nothing but the truth,” or in standard terms, good precision and good recall. Even if circle B were entirely included in circle C, so that all of the relevant information was returned, circle C would still contain a large amount of irrelevant information that must be waded through to find the relevant.

Clearly what is needed is a next generation technology that delivers high performance in both of these areas.

## the cost of conventional search technology

Conventional search technologies can produce perfect recall on some tasks, that is, they return all of the relevant information. Where they fail, though, is that they also return lots of other irrelevant information. And here is where the true cost lies: since mission-critical operations are generally labor intensive and include the best knowledge workers an enterprise has to offer, the amount of time and effort it takes for a knowledge worker to find the right information is the true ongoing cost of the technology.

The primary reason that conventional search technology fails to return precise results is that these technologies can retrieve only word forms or patterns of word forms divorced from their contextual meaning. Users of Internet search engines are aware that a query for “AIDS virus,” i.e. in quotes, is different than the query *AIDS virus* without the quotes: the former is a search for the exact phrase, while the latter is a search for a document containing both words, or documents that contain one or the other word. But a problem arises with documents that contain sentences such as *New software aids virus detection*,

which would be retrieved when the user queried for medical information. So even though the search was for the exact phrase, the conventional technology can not recognize the meaning of the phrase in that context.

Microlanguage has demonstrated<sup>2</sup> that conventional search technology wastes an average of 80% of a knowledge worker's time by returning irrelevant information. Assuming for argument's sake that conventional search technology can routinely deliver all of the relevant information, the relevant information is still only 20% of the total returned information, and mixed in with 80% irrelevant information. Thus only 20% of the knowledge worker's time is productive using conventional search technology.

## thematix technology: the long-term low-cost solution

Thematix is Microlanguage's intelligent language processing technology. It goes beyond simply finding words or patterns of words in a document by being able to recognize the context that words are used in and thus returning more relevant information and less irrelevant information: **the whole truth and nothing but the truth**. A test conducted using third-party materials<sup>3</sup> showed that Thematix technology far outperforms conventional search technology, returning an average of 97.5% precision as against an average of 20% precision returned by the conventional technology.

Thematix is able to return such high precision because it is designed to overcome some of the major difficulties inherent in language<sup>4</sup>. Thematix technology is really two related technologies that work together to distinguish irrelevant information from relevant:

- The **parsing engine** breaks down the input into its constituent words and punctuation.
- The **grammarbase** is a repository of knowledge that gives further instruction to the parsing engine.

Any knowledge can be programmed into a grammarbase so that what is considered relevant and irrelevant is defined according to the needs of the user. In this way only the most relevant information is returned, and the knowledge worker spends time putting that information to use instead of looking for it.

---

<sup>2</sup> Microlanguage Whitepaper, *Intelligent Language Processing Compared With Conventional Search Technology*, available at <http://microlanguage.com>.

<sup>3</sup> Ibid.

<sup>4</sup> For details, see the Microlanguage Whitepaper *Overcoming The Inherent Difficulties In Language With Intelligent Language Processing*.

## conclusion

It is finally possible to reclaim the \$2 billion dollars a year spent sorting through irrelevant information. The true cost of any knowledge management application is not the one-time costs paid up front for the application, but the ongoing costs of paying knowledge workers to use it. Microlanguage has shown that the conventional search technology at the heart of knowledge management applications accounts for the \$2 billion yearly productivity loss by returning an average of 80% irrelevant information that must be sorted through. Microlanguage's Thematix technology returns an average of 97.5% relevant information, and so much more of the money paid to the knowledge worker's goes toward the knowledge worker using relevant information instead of sorting through irrelevant information.