
UpSearch

Recuperação de Informação Relevante

White Paper

UpLexis Tecnologia S/S Ltda.



UpLexis Tecnologia Ltda.

Rua Luís Coelho, 340, 10º andar
Cerqueira Cesar, São Paulo - SP
CEP 01309-903, Brasil

UpSearch

Recuperação de Informação Relevante

White Paper

UpLexis Tecnologia S/S Ltda.

Este documento encerra um descritivo técnico-funcional da nova versão do sistema *UpSearch*, produto que consolida nossos esforços no sentido de viabilizar um modelo eficaz de organização e acesso ao conhecimento organizacional, seguindo as principais tendências mundiais no campo da *Recuperação de Informação*. Em adição aos métodos estatísticos convencionais, foi dado ênfase no emprego de recursos linguísticos para o português brasileiro, enfatizando um modelo híbrido mais enriquecedor para o usuário final no processo de busca e descoberta de informação relevante.

Atenciosamente,

João Marcelo A. Arcoverde
UpLexis - Diretor de P&D

Agosto, 2008

Índice

1	A Nova Geração de Sistemas de RI	1
2	<i>UpLexis</i>: Uma Breve Apresentação	3
3	<i>UpSearch</i>: Funcionamento Básico	5
4	Arquitetura	7
4.1	O Indexador	7
4.2	O Tableter	8
4.3	O Buscador	9
5	Principais Características	10
5.1	Escalabilidade	10
5.2	Desempenho	10
5.3	Disponibilidade	11
5.4	Segurança	11
5.5	Integrabilidade	12
5.6	Filtros e Repositórios	12
5.7	Hierarquia de <i>Caches</i>	12
5.8	Independência de Idiomas	13
5.9	Flexibilidade de Busca	13
5.10	Busca Conceitual (<i>semântica</i>)	14
6	Conclusão	14

1 A Nova Geração de Sistemas de RI

As modernas práticas de gestão do conhecimento nas empresas estão cada vez mais dependentes de instrumentos capazes de processar eficientemente informação não-estruturada, ou seja, textos livres em linguagem natural, que constitui a forma mais abundante de registro do conhecimento humano. A sobrecarga de informação desta natureza caracteriza um dos maiores desafios para todos aqueles que buscam respostas imediatas para seus problemas diários, seja na *Web* ou em grandes coleções de textos espalhados nas organizações em forma de *e-mails*, páginas *Web*, documentos publicados em diversos formatos (PDF, DOC, XML, ODF, etc), planilhas, etc. A tecnologia *UpSearch* habilita operações de negócios onde a informação não-estruturada desempenha um papel crítico no processo de tomada de decisão corporativa.

A *Recuperação de Informação* (RI) é um amplo campo multidisciplinar que lida com a estrutura, análise, organização, armazenamento, busca e recuperação de informação relevante. A relevância da informação contida em um documento é caracterizada pelo correspondente grau de importância percebido pelo usuário. Neste universo subjetivo, saber o que se quer buscar e, não menos importante, saber como buscar, tornam-se requisitos imprescindíveis para o sucesso da recuperação eficiente de informação. Os engenhos de busca hoje disponíveis para essas atividades são os principais artefatos funcionais para nos auxiliar quando sabemos o que queremos, isto porque “relevância” é um conceito situacional e dependente de contexto. Portanto, um desafio fundamental para a nova geração de sistemas de RI consiste na sua capacidade de capturar do usuário a forma mais apropriada para articular a sua real necessidade de informação, bem como de analisar toda a coleção de documentos para encontrar aqueles que satisfaçam esta necessidade, retornando instantaneamente o melhor conjunto ordenado de documentos em precisão e cobertura.

Em muitas situações necessitamos explorar e descobrir novos conhecimentos potencialmente úteis e inovadores para satisfazer um determinado objetivo, como por exemplo, o aprimoramento de um processo de negócio ou a obtenção de uma vantagem competitiva. O processo de descoberta de conhecimento em textos, reconhecidamente complexo e subjetivo, é de fundamental importância para as organizações devido ao papel desempenhado nas tomadas de decisão. Nesse processo, a atividade de *Mineração de Textos* é composta por um conjunto de tecnologias emergentes para análise e interpretação de grandes volume de dados não estruturados, essenciais para compor um projeto de sistema de RI “inteli-

gente”, capaz de extrair e correlacionar os conceitos existentes nos textos, minimizando os problemas inerentes da linguagem humana.

A assimilação deste contexto introdutório facilita a compreensão das vantagens proporcionadas pelas funcionalidades do sistema *UpSearch*, que emprega uma combinação diferenciada de tecnologias (métodos estatísticos e linguísticos - este último específico para a língua portuguesa), capazes de extrair e analisar os conceitos existentes nos textos, viabilizando a automatização de operações-chave em processos decisórios e servindo como uma poderosa ferramenta coadjuvante no processo de tomada de decisão nas organizações.

O sistema *UpSearch* está fortemente embasado em diversas pesquisas acadêmicas incorridas durante a última década, com sucessivos experimentos realizados em diversos domínios. Suas tecnologias modulares proporcionam:

- Acurácia;
- Segurança;
- Desempenho;
- Amigabilidade;
- Portabilidade;
- Escalabilidade;
- Integrabilidade;
- Disponibilidade;
- Interoperabilidade;
- Capacidade analítica.

Dentre as principais aplicações que podem se beneficiar do sistema *UpSearch* para busca e integração de informações, destacam-se:

- *Business, Customer, Competitive Intelligence*;
- Gestão do Conhecimento;
- CRM e roteamento de *e-mails*;
- Comércio Eletrônico;
- Crédito, Risco, Fraude e *compliance*
- Portais organizacionais e *intranets*.

Todos os setores da indústria podem usufruir do sistema *UpSearch*, independentemente de segmento, porte ou tamanho, volume ou natureza de informações, plataforma operacional de *hardware* ou *software* adotados. Exemplos: bancos, telecomunicações, escritórios de advocacia, médicos, etc.

A seguir, apresentaremos sucintamente a empresa $U^{\rho}L^{exis}$, seguido do conteúdo técnico-funcional do sistema *UpSearch*.

2 *UpLexis*: Uma Breve Apresentação

A $U^{\rho}L^{exis}$ é uma organização orientada à pesquisa e desenvolvimento de soluções tecnológicas fundamentadas em métodos computacionais relacionados à aquisição, organização e acesso à informação. Seu domínio de atuação abrange o processamento computacional de textos em linguagem natural através de conhecimento fundamentado em áreas interdisciplinares como lingüística, estatística e inteligência artificial.

Nossa empresa está presente no mercado globalizado em diversos projetos relacionados à manipulação de grandes bases de dados textuais, a exemplo de atividades como recuperação e extração de informações, mineração e análise de padrões em textos (provenientes ou não da *Web*). A $U^{\rho}L^{exis}$ possui comprovado reconhecimento por entidades internacionais por notória especialização técnica em algumas das áreas supracitadas.

A $U^{\rho}L^{exis}$ tem como missão a concepção de soluções tecnológicas que habilitem o processo de descoberta de conhecimento potencialmente útil e inovador às operações de negócios corporativas, sobretudo aquelas onde a informação não estruturada (textos-livres) desempenha um papel crítico no exercício da tomada de decisão.

No contexto relacionado ao foco deste *whitepaper*, a $U^{\rho}L^{exis}$ apresenta um ambicioso projeto de Sistema de Recuperação de Informação, capaz de indexar grandes coleções de documentos não-estruturados, possibilitando o usuário recuperar informações relevantes e automatizar operações onde a análise de textos desempenha um papel fundamental no processo decisório.

Vários projetos relacionados à aquisição, extração e processamento de dados foram ou estão sendo conduzidos entre a $U^{\rho}L^{exis}$ e grandes clientes como Bradesco, Banco Fibra,

IDG Brasil, OAB-SP, Equifax, Rede SPC, Cisp, Advocacia Pinheiro Neto, Imprensa Oficial de São Paulo, dentre outros. É uma satisfação poder compartilhar nossa experiência sobre um projeto inovador na área de *Recuperação de Informação*.

Estamos certos de que nosso trabalho beneficiará aqueles processos de negócio cujos objetivos estratégicos estejam alinhados com as potencialidades proporcionadas pelo emprego das novas e emergentes tecnologias da informação.

3 *UpSearch*: Funcionamento Básico

O *UpSearch* é um sistema de armazenamento e recuperação de informações para grandes coleções de documentos, fundamentado nas mais modernas técnicas e ferramentas da atualidade, amplamente utilizadas por sistemas de busca como o *Google*¹, *Yahoo*² e *AllTheWeb*³, por exemplo, porém customizadas para o ambiente corporativo.

Os documentos repassados ao *UpSearch* são transformados em índices de busca - estruturas de dados apropriadas para melhor representar o texto, proporcionando maior desempenho na sua localização. Para recuperar documentos que satisfaçam uma determinada necessidade de informação, o *UpSearch* não abre cada documento da coleção para localizar as informações solicitadas: ele apenas analisa os índices gerados, que representam o conteúdo dos documentos, reconhece aqueles que são relevantes e os apresenta de forma organizada para o usuário no menor tempo de resposta possível, mesmo após ter indexado milhões de documentos. Abaixo, na Figura 1, encontra-se ilustrada a visão geral de um sistema de RI, onde os textos são organizados em índices que, por sua vez, são comparados com as consultas dos usuários, retornando documentos relevantes. O sistema pode proporcionar uma realimentação assistida, ampliando sua capacidade cognitiva através de evidências e contextos capturados pela interação com o usuário.

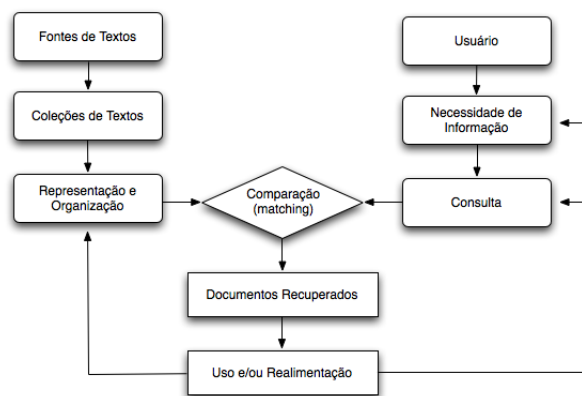


Figura 1: Visão geral de um sistema de RI

¹ www.google.com

² www.yahoo.com

³ www.alltheweb.com

A grande velocidade do *UpSearch* se deve tanto as suas ágeis estruturas de dados, como também a sua arquitetura distribuída em *cluster* - um conjunto de computadores de *hardware* comum que, ao trabalhar coordenadamente, atinge capacidade computacional equivalente ao de supercomputadores, mas a um custo extremamente reduzido. Esta mesma arquitetura é responsável pela escalabilidade do ambiente operacional, i.e., o sistema responde com as mesmas métricas de avaliação, independentemente do tamanho da coleção indexada, podendo variar de algumas centenas de documentos a vários milhões.

A principal linguagem de comunicação homem-máquina ainda é aquela baseada em palavras-chave, através da qual o usuário formula sua necessidade de informação fornecendo ao sistema os termos que melhor descrevem os conceitos almejados, através do uso de operadores *Booleanos* (AND, OR, NOT, etc.), de maneira intuitiva, de fácil manipulação, e que permite ordenar o conjunto de documentos retornados pela consulta em função de algum critério de relevância. Esta ordenação é uma tarefa difícil de ser conduzida, seja pela inabilidade do usuário em saber articular ou mesmo interpretar eficientemente a sua necessidade de informação, seja pela própria natureza da linguagem humana, caracterizada pela ambigüidade, subjetividade e imprecisão. Sendo assim, existe sempre uma distância semântica entre a sua real necessidade e a consulta formulada.

Para minimizar os problemas da linguagem, o sistema *UpSearch* permite utilizar (opcionalmente em tempo de indexação) um conjunto de métodos estatísticos e lingüísticos que habilitam a extração dos conceitos contidos nos textos, organizando-os em função de sua distância semântica. Os fenômenos conhecidos como polissemia⁴ e sinonímia⁵ são minimizados, possibilitando recuperar documentos pertencentes ao mesmo tópico (e, portanto, relevantes), mesmo que contenham diferentes termos para representar os mesmos conceitos. Por exemplo, ao buscar por “automóvel”, espera-se recuperar documentos contendo “veículo” e que não contenham o termo “automóvel”. Desta forma, é possível aumentar a cobertura do sistema sem prejudicar sua precisão, beneficiando a acurácia e proporcionando uma melhor experiência de busca e recuperação para o usuário.

Além da capacidade de localizar eficientemente os documentos relevantes e de poder ser escalável para grandes coleções, o *UpSearch* possui diversas outras vantagens, a exemplo dos níveis de segurança para os documentos - podendo-se configurar permissões de acesso; tolerância a falhas (disponibilidade); interface de fácil operação através de um

⁴ um termo pode representar vários conceitos

⁵ diferentes termos representam o mesmo conceito

ambiente amigável, projetado para contemplar conceitos de usabilidade; capacidade de processar vários tipos de documentos (PDF, XML, HTML, DOC, etc.); portabilidade para automatizar operações onde a análise de textos desempenha um papel central no processo decisório, como por exemplo, atividades de classificação, agrupamento e roteamento de textos.

4 Arquitetura

A arquitetura do *UpSearch* é baseada em *cluster*⁶ (aglomerados), consistindo de um sistema paralelo e distribuído de alta disponibilidade e desempenho.

No *cluster*, as máquinas possuem diferentes papéis para o funcionamento do sistema: haverá um Broker e N servidores. Entretanto, o *cluster* permite ser formado por um único *hardware* com configuração de servidor padrão de mercado, ou pode escalar milhares de servidores de forma distribuída e coordenada, em função do dimensionamento do volume e taxa de crescimento da coleção de documentos.

O Broker é a porta de entrada para as consultas e a inteligência do *cluster*, sendo responsável por receber todas as requisições entrantes e tomar as decisões sobre o que fazer e qual(is) do(s) servidores acionar, distribuindo e coordenando o trabalho entre eles. Estes servidores são todas as máquinas do *back-end* do *cluster*, onde os documentos estão efetivamente armazenados e onde o processamento realmente acontece.

A arquitetura do *UpSearch* é formada por três grandes subsistemas distribuídos no *cluster*, que trabalham de forma integrada: o Indexador, o Tableter e o Buscador. Todos os 3 subsistemas estão parcialmente presentes tanto no Broker quanto nos servidores.

4.1 O Indexador

Este subsistema é responsável por transformar todos os documentos em uma estrutura de dados distribuída, em um processo extremamente eficiente baseado na customização

⁶ conjunto de computadores conectados entre si e que trabalham em conjunto de maneira coordenada, podendo ser vistos sob vários aspectos como se fossem um único super-computador

de uma biblioteca *open-source* denominada Lucene⁷, mundialmente utilizada para indexação de textos e cuja licença de uso⁸ permite sua customização e redistribuição comercial embutidas em aplicações proprietárias.

Em seguida, o índice resultante sofre um processo de distribuição na forma de um *índice global*, conforme a Figura 2. Ao contrário dos índices adotados pela grande maioria dos sistemas existentes no mercado (distribuídos localmente, compartilhados ou replicados), os índices globais conseguem conciliar excelentes valores em todas as características desejadas nesses sistemas: baixo tempo de resposta, alto paralelismo e enorme escalabilidade - sendo assim o mais indicado para grandes coleções.

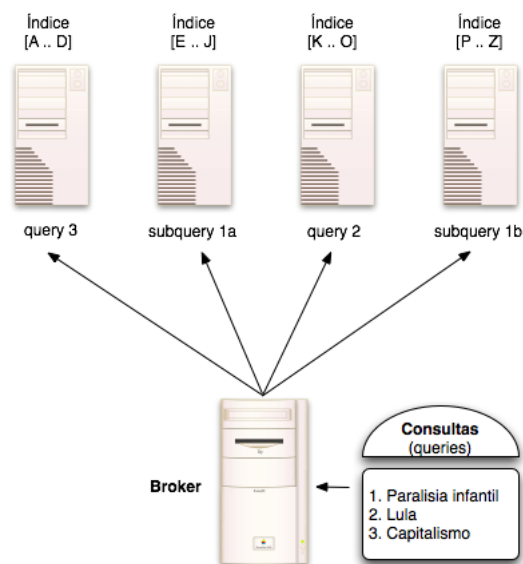


Figura 2: Índice Global distribuído entre múltiplos servidores

4.2 O Tableter

É o subsistema responsável por estruturar os documentos em um sistema lógico de arquivos especiais, projetado para maximizar a organização e recuperação dos documentos diretamente dos dispositivos físicos de armazenamento (discos). Assim, após uma busca

⁷ lucene.apache.org

⁸ www.apache.org/licenses/LICENSE-2.0.html

ser realizada no *UpSearch*, o Tableter é o responsável por exibir para o usuário aqueles documentos mais relevantes.

A grande vantagem do Tableter é o seu desempenho na recuperação dos documentos, superior, inclusive, aos sistemas de bancos de dados relacionais (Oracle, SQL Server, MySQL, etc.). Isso é possível devido a sua extrema especialização, ao contrário dos SGBDs, que são sistemas bastante genéricos. A arquitetura do Tableter faz uso de índices *in-memory*, o que otimiza suas operações.

Da mesma forma que o Indexador, o Tableter também é um subsistema distribuído no *cluster*, dividindo os documentos adicionados ao *UpSearch* de forma homogênea no *cluster*, economizando recursos, equilibrando a carga de trabalho por máquina, favorecendo a escalabilidade.

4.3 O Buscador

Uma vez que os documentos estejam devidamente indexados pelo Indexador e armazenados pelo Tableter, as consultas realizadas no *UpSearch* podem iniciar o processo de busca propriamente dito sobre o índice que representam os documentos.

Graças a organização dos índices criada pelo Indexador, o Buscador não precisa do *cluster* inteiro para realizar cada busca, mas sim apenas uma parte dele. Dessa forma, o Buscador consegue se valer de um altíssimo paralelismo (por *multithreading* e na utilização de várias máquinas) para realizar múltiplas buscas simultaneamente.

Ao final do processo de busca e recuperação, uma página contendo os resultados é apresentada ao usuário, com as respectivas ocorrências agrupadas e ordenadas por algoritmos apropriados, que computam a probabilidade de relevância do documento em relação à consulta, em um processo estatístico extremamente funcional. Neste ponto reside toda a “inteligência” analítica do sistema *UpSearch*, cujo detalhamento técnico foge do propósito deste *whitepaper*. Além disso, diversas informações sobre os documentos são exibidas, como a origem e o tipo de documento, a data e trechos onde o padrão de busca foi encontrado.

5 Principais Características

Abaixo destacam-se as principais características do sistema *UpSearch* que o diferencia em relação a outros engenhos de busca existentes no mercado, a seguir:

5.1 Escalabilidade

Um dos grandes focos do *UpSearch* é a escalabilidade para suportar desde pequenas até grandes coleções de documentos. Todas as técnicas distribuídas utilizadas visam que o sistema se comporte da mesma forma para *clusters* compostos por 1, 10, 1000 ou 100.000 máquinas.

Mais importante ainda, a adição de novas máquinas ao *cluster* constitui uma operação extremamente simples, garantindo a melhora do desempenho do sistema de maneira linear em função do número de servidores adicionados. Assim, ao se dobrar a quantidade de servidores do *cluster*, a capacidade do sistema é praticamente dobrada⁹.

O *UpSearch* também possui uma série de algoritmos capazes de aprimorar a escalabilidade com o crescimento do sistema, devido à redistribuição inteligente da carga do sistema e à redundância de dados, dividindo a carga do sistema por todo o *cluster*.

5.2 Desempenho

O processo de indexação é realizado a uma velocidade em torno de 20Mb por minuto em um Pentium M 1.5 GHz. Esta taxa é consideravelmente ampliada em função da plataforma adotada e da própria natureza e características da coleção de documentos. O *UpSearch* suporta indexação incremental com o mesmo desempenho da indexação em lote. O tamanho final do índice representa cerca de 20-30% do tamanho original da coleção.

⁹ a adição de mais servidores a um sistema distribuído possui uma curva de desempenho que nunca pode se equivar, ou superar, à linear - isto devido a trechos de códigos seqüenciais e aos custos de comunicação (lei de Amdahl)

5.3 Disponibilidade

Sendo um serviço essencial, o *UpSearch* é projetado para ser tolerante à falhas, ou seja, o sistema deve estar em funcionamento com disponibilidade acima de 99,9% do tempo, de tal forma que o ele consiga lidar e corrigir sozinho diversos problemas que surjam durante o seu funcionamento.

Como os dados são replicados entre diferentes servidores do *cluster*, caso o restante do *cluster* perca o contato com uma de seus servidores, o *UpSearch* é capaz de detectar esta perda e então ativar uma das cópias replicadas em outro servidor, mantendo o serviço disponível em um curto intervalo de tempo. Mesmo quedas maiores, como a do serviço *web* da máquina *front-end* do sistema, podem ser contornadas, através da réplica do serviço e redirecionamento de IP.

5.4 Segurança

A segurança de informações é um fator essencial quando documentos e informações importantes estão envolvidos: não é do interesse de nenhuma empresa que funcionários não autorizados ou indivíduos externos tenham acesso a todas as informações. Dessa forma forma, o *UpSearch* dispõe de dois níveis de segurança para sua empresa:

- **Tablets criptografadas**

Todas as informações adicionadas no *UpSearch* são criptografadas, de tal forma que o acesso manual aos arquivos não revela informações ao invasor.

- **Níveis de acesso**

Diversos níveis hierárquicos de acesso podem ser estabelecidos no *UpSearch*, de forma que os usuários autorizados, ao se identificar com suas senhas e estabelecer uma sessão no sistema, apenas poderão acessar àquelas informações e documentos que o seu nível de usuário permite. Poderão, inclusive, saber que existem documentos que satisfazem suas consultas mas que não podem ser recuperados devido aos níveis de permissão, se o administrador do sistema habilitar esta configuração.

5.5 Integrabilidade

O *UpSearch* é capaz de se integrar perfeitamente com diversos outros sistemas já existentes na empresa, aceitando consultas formuladas automaticamente e devolvendo os *templates* de resposta em formato *XML*, amplamente utilizado como protocolo de comunicação máquina-máquina. Além disso, pode funcionar de base para outros sistemas internos, apenas utilizando seus serviços via APIs.

As facilidades do *UpSearch* auxiliam às empresas a desenvolverem seus próprios *softwares* customizados, baseados no funcionamento do *UpSearch*. Qualquer linguagem de programação (incluindo *scripts*) é capaz de utilizar os recursos do *UpSearch*, seja através de comunicação via *sockets*, *middlewares* (CORBA ou RMI) ou mesmo *Web services*.

5.6 Filtros e Repositórios

O *UpSearch* é capaz de indexar uma infinidade de formatos de documentos, bem como obtê-los de repositórios localizados em outras fontes e ainda indexar o conteúdo de bases de dados.

- Formatos de arquivos: XML, PDF, Microsoft Word e OpenOffice, dentre outros;
- Diretórios compartilhados: Windows e Linux/Unix;
- Fontes na internet: via HTTP, HTTPS, FTP e SFTP, dentre outros;
- Bases de dados: MySQL, PostgreSQL, Firebird/Interbase e Oracle, dentre outras.

5.7 Hierarquia de *Caches*

Os *caches* são um clássico recurso da Ciência da Computação para armazenamento temporário de informações previamente utilizadas ou computadas, para que acessos futuros tenham alto desempenho sem necessitar de novas computações, além de reduzir a carga de trabalho do sistema como um todo. No *UpSearch* não existe um único *cache*, mas sim múltiplos *caches* de 3 diferentes tipos:

- ***Result Caches***

Armazenam resultados completos de buscas no servidor *Web* do sistema e os resultados parciais no Broker, acelerando ainda mais essas buscas. Os *Results Caches* podem ser regulados para fazer *cache* das N buscas mais recentes, das N buscas mais freqüentemente realizadas ou ainda um misto dos dois;

- ***List Caches***

Estão localizados em todos os servidores, e decidem quais estruturas mais importantes do índice devem ser mantidas na memória principal quando a parcela do índice armazenado naquele servidor é maior do que a sua memória principal. Esses *caches* reduzem consideravelmente a utilização do disco rígido nos processamentos;

- ***Projection Caches***

São pré-computações de listas de ocorrência conjunta de termos mais freqüentes (com muitas ocorrências) nos documentos indexados, e estão presentes tanto no Broker quanto nos servidores.

5.8 Independência de Idiomas

Em fase de busca, o sistema *UpSearch* trabalha com padrões de símbolos destituídos de valor semântico, diferentemente da fase de indexação habilitada (opcionalmente) para a busca conceitual. Isto torna o sistema completamente independente de idioma. Pode-se indexar e buscar documentos em qualquer língua.

5.9 Flexibilidade de Busca

Os resultados da busca são ordenados em função da relevância dos documentos em relação aos termos da consulta, através de algoritmos de alta eficiência e acurácia. As consultas podem ser formuladas através de expressões *Booleanas*, frases exatas, utilizando-se um poderoso conjunto de operadores lógicos, inclusive permitindo-se expressar proximidade entre termos e realizar busca por aproximação, trazendo documentos com alto índice de vulnerabilidade a erros tipográficos.

5.10 Busca Conceitual (*semântica*)

O sistema *UpSearch* pode ser configurado para utilizar métodos estatísticos e lingüísticos para identificar correlações entre os textos. As conexões ente os termos que compõem os textos são compreendidas através do cálculo das probabilidades de associação entre eles, computando-se assim as respectivas distâncias semânticas entre os conceitos expressos pelos principais termos dos documentos: os sintagmas nominais¹⁰.

Na prática, é permitido ao usuário buscar por um determinado conjunto de termos e retornar documentos que não necessariamente os contenham, mas cujo seu conteúdo possua uma forte relação semântica com os termos que foram informados na consulta. Muito provavelmente os documentos recuperados se referem a um mesmo assunto ou pertencem a um mesmo tópico. Esse artifício eleva a cobertura¹¹ do sistema sem prejuízo substancial da precisão¹², muito embora demande um tempo adicional na fase de indexação e tem que ser refeito periodicamente à medida que o tamanho da coleção cresce ao longo do tempo.

6 Conclusão

A tecnologia *UpSearch* constitui uma alternativa eficiente, escalável e sobretudo segura dentre as soluções existentes para o gerenciamento de documentos não-estruturados, de fácil adaptação em ambientes heterogêneos, a um custo extremamente justificativo.

As novas gerações de sistemas de Recuperação de Informação tendem a ser mais inteligentes e amigáveis, permitindo com que seus usuários lidem cada vez melhor com o fenômeno da sobrecarga de informação, auxiliando-os no processo de busca e descoberta de conhecimento potencialmente útil e inovador.

A *U^ρL^{exis}* avança continuamente rumo ao progresso de suas potencialidades tecnológicas, fornecendo meios para que a máquina, ao ampliar suas faculdades cognitivas, possa exercer um papel cada vez mais importante como instrumento de apoio ao processo de decisão nas operações corporativas.

¹⁰ correspondem às estruturas sintáticas de maior carga semântica e poder informativo

¹¹ número de documentos relevantes recuperados em relação ao total de relevantes existente

¹² número de documentos relevantes recuperados em relação ao total de documentos recuperados