NovaIntell – Projecto de Text Mining para a língua portuguesa numa empresa de Gestão de Informação e Conhecimento

por

Pedro Gonçalo Jorge Rolim

Trabalho de projecto apresentado como requisito
parcial para obtenção do grau de

Mestre em Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão de Informação
da
Universidade Nova de Lisboa

NovaIntell – Projecto de Text Mining para a língua portuguesa numa empresa de Gestão de Informação e Conhecimento

Trabalho de projecto orientado

por

Professor Doutor Miguel de Castro Neto

Instituto Superior de Estatística e Gestão de Informação

da

Universidade Nova de Lisboa

Julho 2011

Agradecimentos

Pretendo agradecer à equipa da Manchete todo o apoio dado na realização deste projecto, em particular à Dra. Fátima Rebelo por me ter confiado o projecto NovaIntell e à Dra. Sílvia Gaspar pela sua disponibilidade.

Gostaria igualmente de agradecer todo o apoio dado pelo Prof. Dr. Miguel de Castro Neto, não só na orientação do meu trabalho mas também pelo empenho dado desde a primeira hora ao projecto NovaIntell.

Finalmente, um agradecimento muito especial à minha família cujo apoio incondicional foi indispensável ao longo de todo este percurso.

Resumo

O constante aumento da informação escrita disponível cria um crescente problema de capacidade de análise dos conteúdos textuais. Ao contrário da informação numérica ou categorizada sobre a qual está disseminada a utilização de processos analíticos com vista à extracção de conhecimento, a informação textual é não estruturada. Nesta medida, a sua análise seja mais complexa, menos eficiente e insuficiente tendo em consideração o volume de informação a tratar. Por outro lado, no momento do lançamento do projecto Novaintell, as soluções analíticas de texto que contemplem processos específicos para a língua portuguesa não se encontravam suficientemente desenvolvidas para que fossem adoptadas como alternativas viáveis e eficazes ao tratamento dos conteúdos textuais realizado de forma manual pelos analistas de informação.

Foi com o objectivo de solucionar o problema do tratamento da informação textual que se desenvolveu o projecto NovaIntell, resultado da constituição de um consórcio co-promovido pela empresa de gestão de informação Manchete e pelo Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, apoiado por fundos comunitários no âmbito do Quadro de Referência Estratégico Nacional (QREN). Neste contexto, os trabalhos efectuados ao longo do projecto NovaIntell resultaram no desenvolvimento de um módulo linguístico capaz de analisar o conteúdo de documentos escritos na língua portuguesa, com vista à extracção de conhecimento accionável com base na informação produzida pelos órgãos de comunicação social portugueses.

Paralelamente, o projecto NovaIntell visou igualmente a promoção de uma nova forma de obtenção de conhecimento e respectiva aplicação no mercado de uma área de conhecimento emergente, text mining, que tem vindo a despertar a atenção dos mais diversos sectores de actividade económica, na medida que a exploração e a criação de valor com base em repositórios de conhecimento não estruturado são uns dos principais desafios que se colocam às organizações que operam na esfera da sociedade de informação e do conhecimento.

Palavras-chave: Text Mining, Gestão de Informação, Inteligência Competitiva

Abstract

The increasing amount of textual information available has created a problem regarding the incapacity to analyze all its contents. Unlike numerical or categorized information which the use of analytical processes for the extraction of knowledge is disseminated on, textual information is unstructured. To this extent, their analysis tends to be more complex, inefficient and inadequate taking into account the volume of information to handle. On the other hand when the NovaIntell project has started, textual analytics solutions that include processes to analyze documents written in Portuguese language were not sufficiently mature in order to be adopted by information analysts as viable and effective alternatives to manage textual contents in a manually manner.

NovaIntell project was created to mitigate the problems related with the incapacity to handle textual information through an efficient and proactive approach. This project has born with the creation of a co-promoted consortium with the information management company Manchete and the Higher Institute of Statistical and Management Information of Universidade Nova de Lisboa, sponsored by European Community funds under the National Strategic Reference Framework (QREN). In this context, the work carried out throughout the project NovaIntell resulted in the development of a linguistic module capable of analyzing the content of documents written in Portuguese, with a view to extracting actionable knowledge from the information produced by the Portuguese media.

In parallel, the NovaIntell project also sought to promote text mining as new way of acquiring knowledge and its application in the market for an emerging area of knowledge, which has come to the attention of many sectors of economic activity, to the extent that exploitation and the creation of value based on unstructured repositories of knowledge are one of the main challenges faced by organizations operating in the sphere of information society and knowledge.

Keywords: Text Mining, Information Management, Competitive Intelligence

Lista de Acrónimos e Abreviaturas

ACI – Academy of Competitive Intelligence

AdI – Agência da Inovação

AMEC – Association for Measurement and Evaluation of Communication

AR – Análise de Reputação

ATS – Advanced Technology and Systems

BI – Business Intelligence

CAPSI – Conferência da Associação Portuguesa de Sistemas de Informação

CRM – Customer Relationship Management

FIBEP – Fédération Internationale des Bureauxs d'Extraits de Presse

I&D – Investigação e Desenvolvimento

I&DT – Investigação e Desenvolvimento Tecnológico

IC – Inteligência Competitiva

IDC – International Data Corporation

ISEGI-UNL – Instituto Superior de Estatística e Gestão da Informação da Universidade Nova de Lisboa

KBP – Knowledge Based Portal

KDD – Knowledge Discovery in Databases

MLTT – Multilingual Theory and Technology

OCR – Optical Character Recognition

POR Lisboa – Programa Operacional Regional de Lisboa

POS – Part-of-Speech

QREN – Quadro de Referência Estratégico Nacional

SCIP – Strategic and Competitive Intelligence Professionals

SCT – Sistema Científico e Tecnológico

SIIA – Software & Information Industry Association

SEW – Strategic Early Warning

SMS – Short Message Service

TIC – Tecnologias de Informação e Comunicação

URL – Uniform Resource Locator

XeLDA – Xerox Linguistic Development Architecture

XML – Extensible Markup Language

XRCE – Xerox Research Center Europe

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

Desde o início da História que o Homem utiliza a escrita para preservar e transmitir o seu saber. Ao longo dos tempos, o Homem foi criando regras sobre a forma de registar e preservar o conhecimento, nomeadamente com a definição de alfabetos que permitiram sistematizar a escrita em texto. O texto passou então a ser o processo formal de transmissão de conhecimento entre pessoas.

Mais recentemente, com o advento da sociedade de informação, a quantidade de informação textual gerada e ao dispor de cada um de nós tem crescido de forma exponencial. Com este incremento da informação disponível, deixou de ser viável ou sequer possível, continuar simplesmente a ler os documentos para permanecer actualizado relativamente a qualquer tema do nosso interesse ou área de conhecimento. Passou portanto, a existir um fosso cada vez maior e em permanente alargamento entre a quantidade de dados e informação produzidos e a capacidade de os ler, assimilar e de os converter em conhecimento. Por outro lado, as soluções de gestão de informação oferecidas pelas tecnologias de informação e comunicação (TIC) estão vocacionadas para o tratamento de dados estruturados, não sendo por enquanto eficazes na gestão de conteúdos textuais.

Esta incapacidade de gerir todos os dados e informação disponíveis com as ferramentas existentes criou uma vulnerabilidade às pessoas e organizações, na medida que enfrentam grandes dificuldades em tirar partido deste conhecimento valioso para o desempenho das suas actividades. Foi no sentido de mitigar este *deficit* de conhecimento que se desenvolveu este projecto de text mining, consubstanciado na criação do primeiro módulo linguístico de estruturação e análise dos textos escritos na língua portuguesa, numa óptica da gestão de informação e conhecimento.

1.1. Enquadramento

Este trabalho de projecto apresenta uma síntese das actividades realizadas na criação de um módulo analítico com vista à estruturação e análise sistematizada dos textos escritos em língua portuguesa. O trabalho descrito resultou do desenvolvimento de um projecto em co-promoção estabelecido entre a empresa Manchete – Gestão de Informação, SA e o Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa (ISEGI-UNL).

O projecto NovaIntell visou dotar em primeiro lugar a Manchete e posteriormente o mercado, de uma tecnologia de análise textual anteriormente inexistente e consolidar o ISEGI-UNL como uma referência nacional e internacional nesta área de grande relevância para o futuro da gestão de informação e conhecimento como é o caso do text mining.

A importância estratégica do projecto NovaIntell no contexto da economia nacional, bem como o valor académico associado foram avalisados pela Agência da Inovação (AdI) ao conceder apoios comunitários no âmbito do Programa Operacional Regional de Lisboa (POR Lisboa) integrados no Quadro Referência Estratégico Nacional (QREN).

1.2. Motivação

O projecto NovaIntell teve como propósito a utilização das técnicas e processos de text mining que permitiram não só melhorar a forma como a Manchete dirigia os processos internos de gestão de informação mas valorizar simultaneamente, um dos seus maiores activos que é o repositório de dados não estruturados que possui um acervo documental com mais de quatro milhões de notícias em formato digital.

Com a aquisição de competências de aplicação de processos analíticos sobre texto, a Manchete passou a poder diversificar o universo de conteúdos noticiosos trabalhados, dispondo de maior capacidade para processar de forma sistematizada

os conteúdos em formato digital que proliferam na internet, nomeadamente jornais online, comentários às notícias feitos pelos leitores, blogues, fóruns e redes sociais, etc. Consequentemente, a Manchete passou a ter ao seu dispor um conjunto de ferramentas eficientes na elaboração de estudos de Strategic Early Warning (SEW), Inteligência Competitiva (IC) e Análise de Reputação (AR).

Por outro lado, a perspectiva de obter capacidades e experiência no desenvolvimento de soluções de TIC inovadoras, como são o caso dos processos de text mining, resultou num factor adicional de motivação para a execução do projecto NovaIntell.

1.3. Objectivos

No intuito de criar um método automatizado de obtenção de conhecimento baseado nas notícias produzidas pelos órgãos de comunicação social portugueses processadas pela Manchete, o projecto NovaIntell assentou na criação de um processo computacional composto por regras linguísticas e dicionários de termos específicos a cada área de interesse a aplicar sobre textos escritos na língua portuguesa.

Nesse sentido, o objectivo principal do projecto residiu no desenvolvimento do módulo linguístico capaz de identificar entidades tão diversas como nome de pessoas, empresas, funções, quantidades, etc., e identificar a relação existente entre elas. Pretendeu-se que este módulo linguístico contemplasse a possibilidade de alicerçar componentes adicionais, ou seja, capacidade de gerar uma base sobre a qual puderam assentar os desenvolvimentos de módulos linguísticos complementares, focados na obtenção de outro tipo de resultados. A categorização de documentos, a análise de sentimentos ou a realização de investigações científicas, são alguns exemplos dessas componentes.

O segundo objectivo do projecto NovaIntell relacionou-se com a definição de modelos de IC. Estes modelos de descoberta e valorização do conhecimento assentaram na aprendizagem e correcta utilização do sistema com vista à

percepção de elementos valiosos no contexto de uma determinada análise ou investigação.

A divulgação dos resultados alcançados e a promoção das potencialidades do text mining, enquanto ferramenta de estruturação de dados e descoberta de conhecimento sobre repositórios de dados textuais, estabeleceram o terceiro objectivo do projecto NovaIntell.

1.4. Organização da Tese

O primeiro capítulo tem carácter introdutório, no qual é feita uma breve apresentação do projecto NovaIntell, o seu enquadramento, a motivação e objectivos propostos.

No segundo capítulo é feita uma abordagem ao text mining e à estruturação de dados e informação. Neste capítulo são levantadas questões sobre a necessidade de utilização de processos analíticos de texto, sendo feita a revisão bibliográfica sobre dados, informação e conhecimento, estruturação de dados e informação, text mining e aos seus processos.

O terceiro capítulo descreve o projecto NovaIntell e qual o seu enquadramento com o contexto de negócio da Manchete. Neste capítulo é também feita uma caracterização pormenorizada das etapas levadas a cabo durante a realização do projecto.

No quarto capítulo, são apresentadas cada uma das etapas que compõem o processo linguístico adoptado ao longo do projecto NovaIntell e descritos os desenvolvimentos levados a curso durante a construção do *skill cartridge*, explicando a lógica subjacente à criação de dicionários de termos e construção de regras linguísticas desenvolvidas.

No quinto capítulo são apresentados os resultados obtidos com o desenvolvimento do *skill cartridge*, ilustrados com um exemplo de análise no contexto da IC.

O sexto capítulo apresenta as conclusões extraídas do desenvolvimento do projecto NovaIntell. São também apresentadas as perspectivas de desenvolvimentos futuros abertas com a criação do *skill cartridge*.

Os capítulos finais incluem a bibliografia consultada e anexos.

2. Text Mining e Estruturação de Dados e Informação

Algumas das expressões que são recorrentemente utilizadas para definir a sociedade actual e o enquadramento desta no mundo são «*aldeia global*», «*era digital*», «*sociedade do conhecimento*» ou «*globalização*». Independentemente do contexto particular em que cada uma das expressões pode ser usada, os elementos comuns a todas elas relacionam-se com o volume e a facilidade de acesso à informação disponível. O desenvolvimento das TIC tem contribuído para a informatização dos estados, das empresas e dos indivíduos, fazendo com que a quantidade de informação acessível a todos seja cada vez mais vasta e em permanente expansão.

Em resultado destas circunstâncias, as organizações tendem a adaptar-se a esta nova forma de interacção com o mundo. Por não viverem isoladas, as empresas e indivíduos sentem uma necessidade cada vez mais premente de se prepararem para o embate com a concorrência de modo a sobreviverem num mercado cada vez mais globalizado e competitivo. Foi neste contexto que emergiu a IC, processo de suporte à decisão baseado na monitorização do ambiente circundante e análise dos novos factos associados às questões internas. A IC efectiva é um processo contínuo, que envolve a recolha de informação, respeitando os aspectos legais e éticos, análise, que não evita conclusões indesejadas e difusão controlada de *intelligence* accionável aos decisores ("About SCIP").

Por conseguinte, a vida de cada organização assenta em eixos de competitividade específicos do seu posicionamento no mercado que determinam o seu sucesso. Para a generalidade das áreas de actividade, as influências macro ambientais cujo acompanhamento é particularmente importante fazer são a concorrência, os reguladores, os mercados financeiros e a opinião pública (Sullivan, 2001). A esta lista de influenciadores é possível acrescentar os clientes, os fornecedores e outras forças de pressão (Rebelo, 2009).

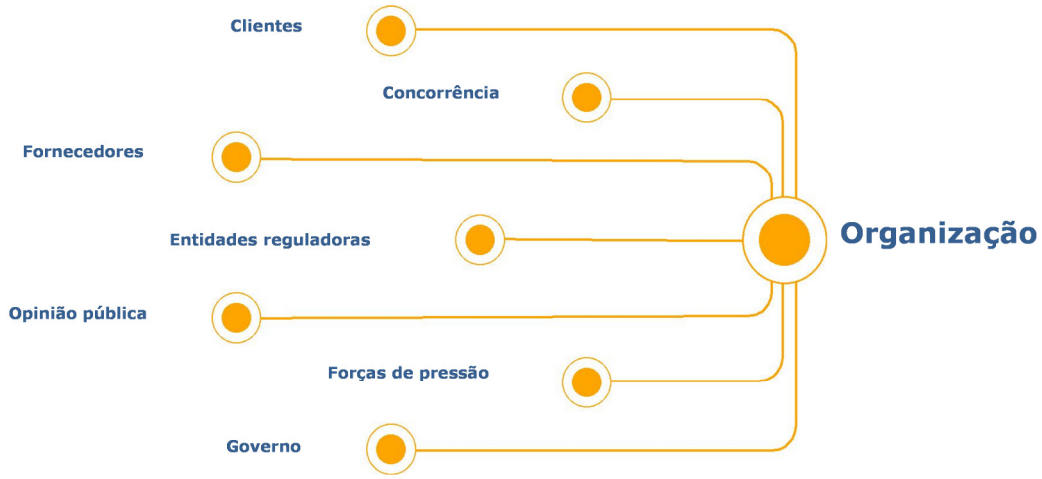

Figura 1 – Eixos de Competitividade

É da dinâmica destes eixos de competitividade que advém o êxito ou o fracasso, não apenas de uma decisão, como até de toda uma estratégia ou, no limite, da sobrevivência de uma organização. A monitorização e controlo do ambiente competitivo através do acompanhamento dos diferentes *stakeholders* asseguram uma vigilância permanente e sistematizada de todos os eixos de competitividade da organização numa perspectiva única e personalizada. Deste modo, sempre que se verifique alguma alteração em qualquer um destes eixos que possa condicionar a actuação da empresa, esta seja a primeira a saber (Rebelo, 2009).

Apesar destes processos de alerta estratégica precoce ou *Strategic Early Warning* (SEW) ajudarem os gestores a compreender as mudanças no ambiente em torno das suas organizações mais rapidamente que a concorrência, permitindo-lhes agir pró-activamente (Ansoff, 1980; Luken & Baisch, 1998), não garantem por si só a melhor decisão em cada momento. Contudo, proporcionam ao decisor que cada decisão tomada seja devidamente sustentada e que teve em conta toda a informação e conhecimento disponíveis à data.

Na perspectiva de satisfazer esta necessidade de obtenção de elementos relativamente ao ambiente empresarial envolvente surgem cada vez mais empresas especializadas na recolha, tratamento e divulgação de informação sobre cada sector de actividade em particular. Como especialistas nos respectivos

sectores de actividade, essas empresas têm necessidade de recolher, manipular e, geralmente, produzir grandes volumes de informação, como é o caso da *Reuters* que em 2005, publicava diariamente o equivalente a três bíblias de notícias somente relacionadas com o sector económico e financeiro (Zanasi, 2005b).

Estes factores suscitam a questão da capacidade de tratamento da informação textual de forma eficiente e sistemática por parte das organizações que dela dispõem:

> *"Os documentos em texto são expressões da língua natural sendo, portanto facilmente perceptível pelos humanos. Contudo, de modo a permitir que computadores nos auxiliem a processar a informação textual, temos de proporcionar a correcta representação do texto. Geralmente envolve o processamento linguístico e estatístico dos documentos através de técnicas que foram desenvolvidas para diferentes tipos de situações."* (Milic-Frayling, 2005, p. 1)

Neste contexto, torna-se pertinente colocar a seguinte questão: Serão as organizações capazes de obter elementos pertinentes dos conteúdos textuais que dispõem ou têm acesso, num período de tempo válido e com os recursos técnicos e financeiros admissíveis?

Nas próximas secções são apresentados alguns dos conceitos necessários para dar resposta à questão anterior.

2.1. Dados, Informação e Conhecimento

As definições mais comuns de dados, informação e conhecimento no contexto das TIC são baseadas na interpretação dada por Russel Ackoff. Diversos autores (Bellinger, Castro, & Mills; Nurnberger, Seising, & Wenzel, 2009) referem que Ackoff definiu dados, informação e conhecimento da seguinte forma:

- Dados: uma sequência de símbolos sem significado;
- Informação: resulta do processamento de dados que permitem dar resposta às questões "quem", "o quê", "onde" e "quando";
- Conhecimento: resulta do relacionamento de diversa informação, sendo capaz de dar resposta à questão "como".

No contexto da Gestão do Conhecimento, diversos autores apresentam múltiplas definições de dados, informação e conhecimento. A tabela 1 apresenta uma síntese de algumas dessas definições, numa recolha efectuada por Stenmark (2002).

Autor/Autores	Dados	Informação	Conhecimento
Wiig, 1993		Factos organizados que descrevem uma situação ou condição	Verdades ou crenças, perspectivas e conceitos, julgamentos e expectativas, metodologias e *know how*
Nonaka e Takeuchi, 1995		Sequência de mensagens com sentido	Crenças e compromissos resultantes das mensagens
Spek e Spijkervet, 1997	Símbolos não interpretados	Dados com significado	Capacidade de alcançar entendimento
Davenport, 1997	Observações simples	Dados com relevância e significado	Informação valorizada pela mente humana
Davenport and Pursak, 1998	Conjunto de factos discretos	Mensagem capaz de mudar a percepção do receptor	Experiências, valores, factos e informação contextualizada

Continua na página seguinte

Continuação da página anterior

Quigley e Debons, 1999	Texto que não responde a questões de um problema em particular	Texto que responde às questões quem, quando, o quê e onde	Texto que responde às questões porquê e como
Choo *et al.*, 2000	Factos e mensagens	Dados revestidos de sentido	Crenças fundamentadas

Tabela 1 – Definições de Dados, Informação e Conhecimento (Stenmark, 2002)

Por seu turno, Courtney (2001) e Tuomi (1999) sintetizam da seguinte forma o que designam como abordagem convencional dos conceitos de dados, informação e conhecimento: Dados são factos em bruto, isolados ou observações simples; informação resulta da estruturação dos dados, da sua contextualização ou de alguma interpretação humana; conhecimento resulta da interpretação da informação accionável, ou seja, da capacidade de agir em função da informação obtida.

Todas estas definições têm subjacente uma relação directa e hierárquica entre dados e informação e entre informação e conhecimento. Esta relação hierárquica é comummente representada pela *Pirâmide do Conhecimento* (figura 2), sendo os dados a sua base, a informação assente sobre os dados e sobre a qual se alicerça o conhecimento, surgindo como vértice superior da figura.

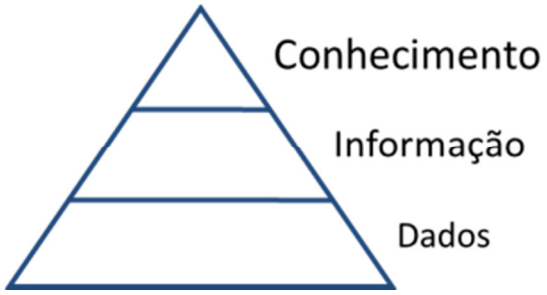

Figura 2 – Pirâmide do Conhecimento em três níveis

No entanto, autores como Carlisle (2007), Tuomi (1999) e Hey (2004) identificaram situações em que a hierarquia tradicional de dados, informação e conhecimento representada pela *Pirâmide do Conhecimento* não é aplicável (Hey, 2004), chegando mesmo a criticar a sua validade (Carlisle, 2007; Tuomi, 1999).

No âmbito do projecto NovaIntell adoptou-se a hierarquia piramidal de dados, informação e conhecimento por ser a que melhor se aplica aos conceitos subjacentes a este projecto de text mining.

As palavras, algarismos e sinais que compõem o texto, arrumadas nas suas categorias gramaticais constituem a base da pirâmide, ou seja, os dados. A agregação e contextualização dos dados permitem obter informação. A análise e interpretação dos vários pedaços de informação permitem alcançar o conhecimento.

O exemplo apresentado da figura 3 permite visualizar a hierarquia piramidal existente entre dados, informação e conhecimento. Na base da pirâmide encontram-se os dados, na forma de um conjunto de substantivos, preposições, verbos e números. Com a agregação e contextualização dos vários dados resultou a captação de informação relativamente aos sujeitos, Oracle e Sun Microsystems, uma acção, compra, e um valor monetário, 5,6 mil milhões de euros. A interpretação dos dados permitiu descobrir uma relação entre as empresas. Obteve-se o conhecimento acerca da aquisição da Sun Microsystems por parte da Oracle no valor de 5,6 mil milhões de euros.

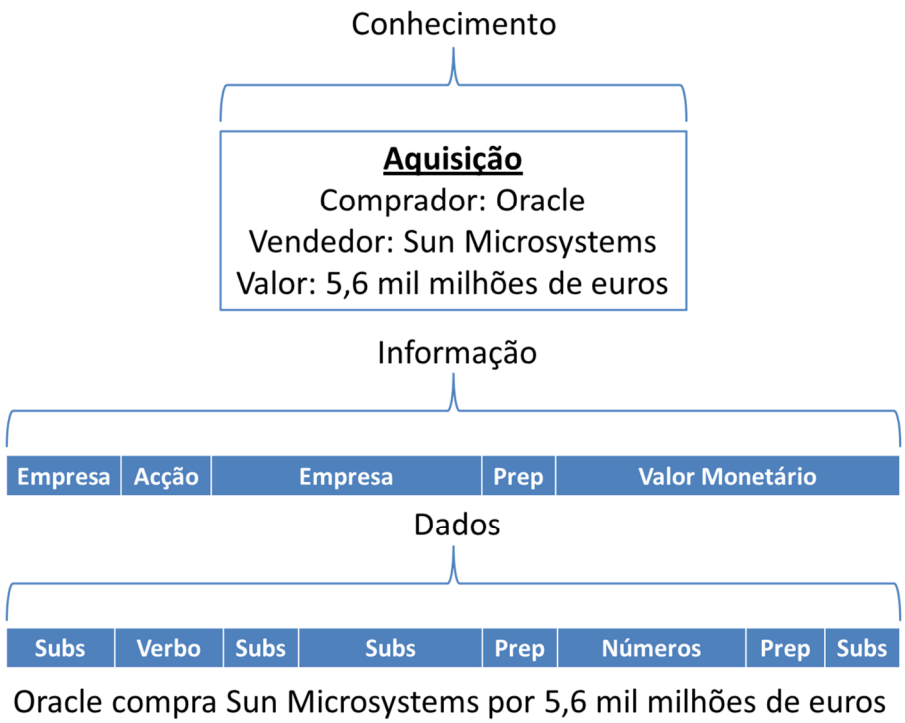

Figura 3 – Exemplo de dados, informação, conhecimento e a sua relação hierárquica

Após a definição de dados, informação e conhecimento, é feita de seguida uma abordagem às propriedades dos dados e informação relativamente à sua estruturação.

2.2. Estruturação de Dados e Informação

Como foi referido no início deste capítulo, os textos são expressões da língua natural, sendo necessário um tratamento desses conteúdos para que computadores nos auxiliem a processá-los. Esse processo de tratamento dos documentos que permite o processamento computacional dos conteúdos é designado por estruturação. Nesta secção é definido e apresentado o processo de estruturação de dados.

Os dados e, por conseguinte a informação, tendem a ser heterogéneos quanto à sua estrutura. Domenig e Dittrich (1999) apresentam a seguinte classificação de

dados com base na forma como são armazenados e quais as operações que são possíveis de efectuar sobre os mesmos (Domenig & Dittrich, 1999):

- Dados Estruturados: respeitam uma matriz bem definida que distingue a sua composição de outros tipos de dados ou resultam da composição de um único tipo de dados como são, por exemplo, os números inteiros.

  Neste contexto, a matriz é definida à priori, ou seja, antes do armazenamento dos dados, é explícita, ou seja, guardada de forma independente dos dados, é rígida (os dados têm de respeitar sempre a estrutura) e pode ser usada nas pesquisas sobre dados.

- Dados Não Estruturados: não são uma composição simples de algarismos ou letras, ou não respeitam uma matriz bem definida.
- Dados Semiestruturados: obedecem a uma estrutura que não é rígida, ou seja, nem todos os dados respeitam a matriz e/ou a matriz nem sempre é separável dos dados.

O conceito dos dados estruturados respeitarem uma matriz de estruturação, contrariamente aos dados não estruturados, é partilhada por Kosovac, Froese e Vanier (2000). Blumberg e Atre (2003) sintetizam que no contexto dos sistemas de bases de dados relacionais, dados não estruturados são os que não se podem armazenar em linhas e colunas. Nesse mesmo contexto, dados semiestruturados são aqueles que têm associados metadados, elementos sobre os dados que facilmente são armazenados numa base de dados relacional, e dados estruturados são todos os que proliferam numa tradicional base de dados relacional.

Tipicamente, dados e informação não estruturados são os conteúdos presentes nas diversas formas de transmitir texto como são os casos dos livros, jornais, *Short Message Service* (SMS) e emails. Dados ou informação semiestruturada encontra-se em documentos com o formato *Extensible Markup Language* (XML) e dados estruturados são os que se encontram armazenados em linhas e colunas numa base de dados relacional (Li, Ooi, Feng, Wang, & Zhou, 2008; Uys, du Preez, & Uys, 2008).

A estruturação dos conteúdos textuais ocorre quando conteúdos não estruturados são convertidos num formato que permitam a sua análise de forma automática. Este processo de transformação compreende a utilização das seguintes componentes (Zanasi, 2005a):

- Modelo linguístico capaz de separar as frases nas suas componentes (substantivos, adjectivos, verbos, números, datas, etc.);
- Dicionários de termos focados em cada área de conhecimento a contemplar (a título de exemplo, basta considerar a especificidade de termos usados na produção de documentação relacionada com a indústria química);
- Ferramenta que permita definir e acomodar as regras e características linguísticas.

A este processo de transformação, Zanasi (2005a) designa de extracção.

Neste contexto, importa definir extracção de informação[1] e distinguir este conceito do de recuperação de informação[2].

Recuperação de informação é a tarefa de identificação do conjunto de documentos relevantes no contexto de determinada necessidade de informação, de entre um universo de textos mais vasto (Gaizauskas & Wilks, 1998; Milic-Frayling, 2005; Pazienza, 2005). É comum aplicar-se recuperação de informação como sinónimo de recuperação de documentos[3], tendo em consideração uma noção de documento mais vasta (Milic-Frayling, 2005).

Por sua vez, extracção de informação é o processo de recolha automática de determinados tipos de informação pré-definida a partir de textos escritos em linguagem natural. Pode assumir-se como sendo a actividade de obtenção ou

[1] tradução de Information Extraction

[2] tradução de Information Retrieval

[3] tradução de Document Retrieval

descoberta de factos numa fonte de informação estruturada que tem os textos não estruturados como fonte de alimentação (Gaizauskas & Wilks, 1998; Pazienza, 2005).

Porventura, a diferença mais significativa que existe entre a recuperação de informação e a extracção de informação, relaciona-se com o facto da primeira retornar documentos que o utilizador final pretende. A recuperação não implica portanto, que esteja associada à descoberta de factos novos, uma vez que esses factos já se encontram descritos pelo autor do documento, contrariamente ao que pode suceder com extracção de informação, onde novos factos podem emergir (Gaizauskas & Wilks, 1998; Hearst, 1999).

Pazienza (2005) descreve o processo de estruturação com base na extracção de informação em duas grandes fases. A primeira etapa, análise lexical, consiste na associação das palavras e expressões idiomáticas às respectivas categorias gramaticais[4] através da análise morfológica e verificação de dicionários, e no reconhecimento de entidades, como por exemplo datas ou expressões monetárias. A segunda fase relaciona-se com a análise sintáctica com vista à identificação de conjuntos de nomes ou verbos. Considerando que «um sistema de extracção de informação é uma "cascata" de transdutores ou módulos que, a cada passo, acrescentam estrutura e frequentemente perdem informação, desejavelmente irrelevante, através da aplicação de regras que são obtidas manualmente e/ou automaticamente», Hobbs (1993) propôs a seguinte estrutura genérica que considera parcial ou integralmente válida para qualquer sistema de extracção de informação (Hobbs, 1993; Pazienza, 2005):

1. Text Zoner[5], que separa o texto num conjunto de segmentos de texto ou frases. No mínimo, feita a distinção das zonas formatadas das não formatadas;

[4] tradução de Part-Of-Speech (POS)

[5] optou-se por não traduzir a expressão original em inglês

2. Pré processador, que converte segmentos de texto numa sequência de expressões, onde cada uma delas é a sequência de elementos lexicais onde cada elemento lexical é uma palavra com os seus atributos léxicos. Na prática reconhece e normaliza padrões de nomes compostos, datas, horas, valores monetários, etc.;
3. Filtro, que reduz as expressões a um conjunto mais reduzido através da filtragem das palavras ou expressões irrelevantes;
4. Preparser[6], que partindo de sequências de elementos lexicais, tenta identificar de forma fiável, estruturas (ou expressões) comuns de pequena escala existentes. Expressões como "Primeiro Ministro" ou "Presidente da República" são exemplos das composições identificadas;
5. Parser[7], que tem sequência de elementos lexicais ou frases como recursos de entrada e cujo resultado é um conjunto de fragmentos de árvore ou diagrama de estruturação, eventualmente completa. Na prática, visa identificar expressões mais complexas que as identificadas pelo *Preparser*;
6. Combinador de fragmentos, tenta converter um conjunto de fragmentos da árvore ou diagrama de estruturação ou outros, num diagrama de estruturação único para toda a frase ou expressão. Em termos práticos, combina as expressões captadas nas duas fases anteriores com vista a obter expressões completas;
7. Interpretador semântico, gera a estrutura semântica com base na árvore ou diagrama de estruturação. Esta funcionalidade valida as várias hipóteses geradas na fase anterior, excluindo as que não têm sentido;
8. Desambiguador Lexical que converte estruturas semânticas com predicados genéricos ou ambíguos numa estrutura semântica com predicados específicos, isentos de ambiguidades. Situações de palavras que podem ser substantivo ou verbo são exemplos de utilização desta componente;

[6] optou-se por não traduzir a expressão original em inglês

9. Resolução de co-referências ou Processador de discurso, que converte estruturas em árvore em estruturas em rede através da identificação de descritores distintos de uma mesma entidade nas diferentes partes do texto. Por exemplo, as expressões "fabricante alemão" e "Mercedes Benz" que podem surgir em distintas áreas do texto podem ser juntas na expressão "fabricante alemão Mercedes Benz";
10. Gerador de modelos, que cria modelos com base nas estruturas semânticas. Esta componente finaliza o modelo, definindo o modo como os resultados são apresentados.

Como se verá mais adiante, o processo linguístico adoptado na execução do projecto NovaIntell inclui várias componentes ou variações do modelo de Hobbs. Processos como a tokenização, análise morfológica, o identificador de nomes compostos têm correspondência com o Text Zoner, Preparser e interpretador semântico do modelo de Hobbs. Ao longo do capítulo 4, o processo linguístico adoptado e respectivas componentes serão analisadas com mais detalhe, sendo apresentados exemplos dessas componentes.

Após a abordagem aos conceitos de dados, informação e conhecimento, bem como à forma como estes se relacionam entre si e terem sido descritos os aspectos relacionados com a estruturação dos dados, a próxima secção foca-se nos problemas resultantes do elevado volume de conteúdos não estruturados existentes com que as organizações se deparam e/ou terão de lidar num futuro próximo.

2.3. Universo Digital

De um modo geral, as empresas tendem a recorrer aos repositórios de dados próprios, na expectativa de conhecer melhor os seus clientes. Exemplo característico da utilização deste tipo de dados é a adopção de processos de data mining sobre os repositórios de dados estruturados. Não obstante a importância

deste tipo de fonte de dados para a caracterização da actividade ou negócio de uma organização, um estudo da *Merril Lynch* e da *Gartner*, refere que 30% a 40% do tempo de trabalho dos executivos é consumido na gestão de documentação, sendo que mais de 85% da informação corporativa é armazenada de forma não estruturada, ou seja, em texto, video, audio ou imagem (Blumberg & Atre, 2003). No mesmo sentido, diversos autores referem que 80% nos negócios são dirigidos com base em informação não estruturada e 85 a 90 por cento de toda a informação é mantida em formato não estruturado (McKnight, 2005; Plejic, Vujnovic, & Penco, 2008; White, 2005).

Paralelamente, o peso dos dados não estruturados deixou de ser crítico na capacidade de um repositório de dados. Baseado num estudo do Hitachi Group, (Smullen, Tarapore, & Gurumurthi, 2007) referem que os custos de armazenamento de dados passaram de cerca de quatro dólares por Megabyte em 1990 para um valor inferior a um cêntimo de dólar por Megabyte em 2007. Este facto por si só favorece o crescimento dos repositórios de dados dentro das organizações, independentemente da sua função primária e da sua tipologia. Estes autores (Smullen, et al., 2007) citam igualmente um estudo da empresa consultora de *market intelligence*, International Data Corporation (IDC), que apresentou o *Universo Digital* como sendo o conjunto de todos os dados em formato electrónico extraídos, armazenados e replicados por todo o mundo, tendo uma dimensão de 161 exabytes em 2006 para mais de 988 exabytes em 2010[7].

Em 2009, a IDC actualizou as métricas associadas ao *Universo Digital.* Jorge Coimbra, director geral da IDC Portugal apresentou no decorrer de um evento da IDC subordinado ao tema de *Enterprise Content Mangement* (Coimbra, 2009), um novo estudo da empresa no qual se estima que o *Universo Digital* deverá quintuplicar num período de três anos, passando dos 487 exabytes em 2008 para mais de 2400 em 2011 (Webster, 2009). O mesmo estudo destaca ainda os seguintes elementos:

[7] 1 exabyte = 10^{18} bytes

- O peso que a informação não estruturada representa no total do *Universo Digital* ultrapassa os 90%;
- Em 2011, cerca de 70% da informação será criada de forma individualizada;
- A informação gerada excederá a capacidade de armazenamento instalada.

Mais recentemente, em 2010, a IDC rectificou as projecções anteriores relativamente à dimensão do *Universo Digital*, revendo em alta os valores referentes a 2008 e as perspectivas para 2010, cifrando-os em 493,8 e 1.200 exabytes, respectivamente. No mesmo documento Gantz & Reinsel (2010), apresentaram os valores de 2009, 800 exabytes e estimou que em 2020, o *Universo Digital* deverá ascender a 35.000 exabytes, ou seja, 35 zetabytes[8].

Como foi referido, de uma forma geral, as organizações têm-se focado no tratamento e análise de informação estruturada existente dentro nos seus repositórios, não dispondo ainda da capacidade para valorizar os seus conteúdos não estruturados. Desta forma, as organizações não só menosprezam o grosso dos dados que armazenam, dados não estruturados, como não aproveitam todo o potencial oferecido pelos conteúdos do *Universo Digital* ao seu alcance. Nestas condições, as organizações tendem portanto, a não estarem devidamente preparadas para lidar com o choque existente entre a necessidade de serem eficazes na descoberta do conhecimento e a capacidade de lidar com um *Universo Digital* que cresce exponencialmente ano após ano (figura 4). Se "procurar uma agulha num palheiro" é uma tarefa reconhecidamente difícil, a seguinte adaptação da expressão popular retrata da melhor forma o problema descrito: "a procura da agulha torna-se ainda mais difícil num palheiro que não para de crescer".

[8] 1 zetabyte = 10^3 exabytes = 10^{21} bytes

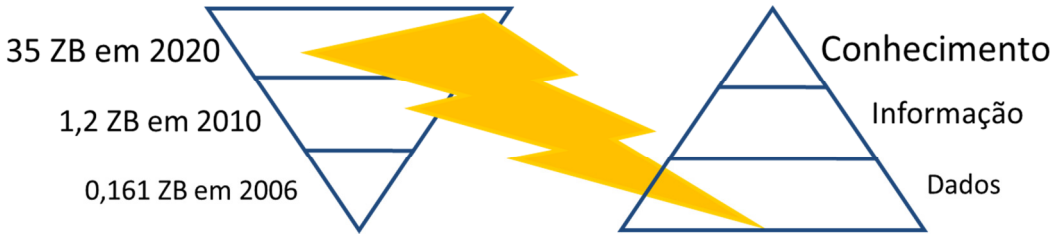

Figura 4 – Representação do choque entre o volume de informação e a descoberta de conhecimento

Na próxima secção é apresentada o conceito de texto mining, através do qual as organizações podem tirar melhor partido dos conteúdos textuais que estão ao seu alcance.

2.4. Text Mining

Text mining também designado por processos analíticos sobre texto ou análise textual[9] abarca, enquanto ferramenta de estruturação e análise de texto, um conjunto de técnicas e metodologias através das quais as organizações podem responder positivamente à questão colocada no início deste capítulo.

Zanasi (2005b) define text mining como sendo uma área multidisciplinar que reúne um conjunto de técnicas como data mining, linguística, aprendizagem máquina, recuperação de informação, reconhecimento de padrões, estatística, bases de dados e visualização de modo a obter uma rápida extracção de informação das grandes bases de dados. Já Prado e Ferneda (2007) sintetizam text mining como sendo a aplicação de métodos e processos computacionais sobre dados textuais com vista a descortinar informação relevante e revelar conhecimento anteriormente desconhecido. Ferrari (2005) descreve text mining como sendo um processo de análise e estruturação de grande volume de

[9] Por se considerar que estas expressões não abarcam todas características intrínsecas do conceito, utiliza-se a expressão original em inglês ao longo do presente documento.

documentos através da aplicação de técnicas computacionais estatísticas e/ou linguísticas.

Kroeze, Matthee e Bothma (2003) sintetizaram uma série de outras definições de text mining, que se apresentam na seguinte tabela:

Autor/Autores	Definição de Text Mining
Chen, 2001	Text mining "executa diversas funções de pesquisa, análise linguística e categorização."
Sullivan, 2000	"Text mining é o estudo e prática de extracção de informação do texto usando os princípios da linguística computacional."
Lucas, 1999/2000	Text mining é "a prospecção de pepitas de novo conhecimento em montanhas de texto, que se tornou acessível a pesquisas computacionais graças à revolução da informação e trabalho na internet." Text mining é "uma forma de examinar um conjunto de documentos e descobrir informação que não reside em nenhum documento em particular."
Hearst, 1999	Text mining, enquanto análise exploratória de dados, é a forma como (desenvolver e) utilizar softwares que assistam o analista a obter informação nova e relevante a partir de um largo conjunto de textos. É um processo parcialmente automático onde o analista participa, interagindo com o sistema.
Biggs, 2000	"Text mining é ideal… para… vislumbrar alterações no mercado ou identificar ideias a seguir."
Albrecht e Merkl, 1998	Text mining é o estabelecer de "relações desconhecidas e insuspeitas sobre bases de dados (de texto)…"
Thuraisingham, 1999 Nasukawa e Nagano, 2001	"Define-se texto mining como sendo data mining sobre texto. Text mining baseia-se na extracção de padrões e associações anteriormente desconhecidas, sobre bases de dados de texto."
Zorn *et al.*, 1999	"Text mining oferece fortes possibilidades de criar conhecimento e relevância a partir de enormes volumes de informação não estruturada disponíveis na internet e intranets corporativas."

Tabela 2 – Síntese de definições de Text Mining (Kroeze, et al., 2003)

Após compilarem estas definições de text mining, assumem que "a essência de text mining é a descoberta ou criação de novo conhecimento a partir de um conjunto de documentos" (Kroeze, et al., 2003). Esta noção de text mining é partilhada por (Delen & Crossland, 2008), que apresentam text mining como sendo "o processo de descoberta de informação nova, anteriormente desconhecida, potencialmente útil a partir de uma variedade de fontes de dados não estruturados, nomeadamente documentos empresariais, referências de clientes, páginas da internet e ficheiros XML".

Num outro contexto, Fayyad, Piatetsky-Shapiro e Smyth (1996) enquadram data mining como sendo uma componente do processo de descoberta de conhecimento em bases de dados[10]. Para estes autores, data mining define-se como sendo "a aplicação de algoritmos específicos com vista à extracção de padrões com base em dados".

Em função das definições anteriores, é possível considerar text mining como sendo uma variante ou especialização dos processos de data mining confinados à informação textual, na medida em que esta metodologia visa igualmente a descoberta de conhecimento "escondido" nos repositórios de dados (Gao, Chang, & Han, 2005; Nasukawa & Nagano, 2001). A tipologia da informação utilizada para alimentar os respectivos processos de análise tende a ser o principal factor de diferenciação entre data mining e text mining. No caso de text mining, a fonte de dados é o texto na forma de expressão natural por contraponto às bases de dados estruturadas usadas em data mining (Delen & Crossland, 2008). Neste aspecto particular e no sentido de consolidar o conceito de text mining e a sua relação com data mining, vale a pena referir que esta área de estudos era inicialmente designada por "*textual data mining*", tendo a expressão "text mining" sido usada pela primeira vez em 1994 por Charles Hout, co-fundador da Temis e actualmente seu *Chief Operating Officer*, num evento do Centro Europeu de Matemática Aplicada da IBM em Paris (Zanasi, 2005b).

[10] tradução de KDD (Knowledge Discovery in Databases)

O interesse crescente por processos de data mining e, mais recentemente de text mining, também se justifica pela possibilidade de aproveitamento dos dados existentes nos repositórios corporativos. Efectivamente, através destes processos torna-se viável a utilização de dados secundários (Mannila, 2000). Aproveita-se, desta forma, dados que foram produzidos pelos diversos sistemas das empresas com intuitos operacionais ou outros, sobre os quais não era expectável à partida, serem usados para análise. As fontes de informação secundária passam a ser um novo recurso para extrair conhecimento e, por conseguinte, gerar valor para a organização. A perspectiva de transformar os dados em informação e de sobre esta desenvolver conhecimento acerca das actividades da empresa e o seu funcionamento ou sobre os seus clientes a partir dos dados armazenados ao longo de anos tornou-se uma tarefa simultaneamente natural e obrigatória para a generalidade das organizações que geram ou manipulam grandes volumes de dados.

Em linha com o que foi referido anteriormente, os processos de text mining para além de proporcionarem nova utilidade à informação corporativa interna, abrem as portas à utilização de novas fontes de conhecimento. À data de hoje, torna-se viável utilizar recursos de informação externos à organização, a maior parte dela gratuita ou com baixos custos de obtenção, para alimentação de processos de IC, SEW, AR (análise das percepções, opiniões e expectativas dos *stakeholders*) (Schanz, 2006) e marketing intelligence (processo de recolha e análise de informação de modo a compreender o mercado, determinar necessidades, preferências e tendências do mercado tanto actuais como futuras e identificar mudanças no ambiente empresarial susceptíveis de alterar a dimensão e natureza do mercado) (Cornish, 1997). De salientar que a utilização de fontes de dados públicas são particularmente úteis nas análises de marketing por serem de fácil e rápido acesso, não dispendiosas e por existirem em grande quantidade (Fleisher, 2008).

2.5. Processos de Text Mining

Uma das mais-valias dos processos de text mining advém da capacidade de transformar a informação textual não estruturada de modo a que esta possa ser integrada numa grande base de conhecimento. Consegue-se desta maneira organizar os dados textuais de forma a tornar possível o desvendar de informação nova e nunca antes encontrada e, a partir daí, obter conhecimento valioso.

Como foi referido anteriormente, os processos de text mining encontram-se intrinsecamente relacionados com processos de recuperação e extracção de informação. Por sua vez, os processos de recuperação e extracção de informação materializam-se na execução de uma ou várias das seguintes componentes (Delen & Crossland, 2008; Pazienza, 2005):

- Sumarização: resumo do documento com vista à poupança de tempo de leitura;
- Categorização: identificação dos principais assuntos do documento para o associar a uma ou múltiplas categorias previamente definidas;
- Clustering: junção dos documentos com características semelhantes, independentemente de uma eventual categorização distinta;
- Relacionamento de entidades: relacionar documentos com base nas entidades extraídas, permitindo vislumbrar relações não óbvias;
- Resposta a questões: melhor resposta a uma questão com base na implementação de regras de reconhecimento de padrões;
- Filtragem: selecção dos documentos pertinentes no contexto de uma determinada análise ou pesquisa;
- Routing: distribuição dos documentos para indivíduos ou grupos com base no seu conteúdo e metadados associados.

Estes processos analíticos deverão ser capazes de criar valor às organizações, podendo para tal, tornar-se complementares aos sistemas de Business Intelligence (BI) que estas, eventualmente, já tenham em funcionamento. Esta complementaridade surge na medida dos tradicionais sistemas de BI serem eficientes na satisfação da necessidade de extrair conhecimento de fontes de dados

estruturados e eminentemente internos, mas não sendo eficazes em obter informação de dados pouco ou nada estruturados, nem estando, de um modo geral, talhados para processar dados exteriores à respectiva organização (Gao, et al., 2005; Sullivan, 2005).

Para obter esta complementaridade de dados num sistema de BI, deve ser criado um repositório de dados textuais, distinto do repositório de dados estruturados. Desta forma, para além do tradicional *data warehouse*, deve ser implementado um *document warehouse*. Sullivan (2001) apresenta quatro atributos que caracterizam um *document warehouse*:

- Não existe uma única estrutura ou tipologia de documentos;
- O repositório é alimentado por documentos oriundos de múltiplas fontes;
- As principais características e conteúdos dos documentos são extraídos, preservados e armazenados automaticamente no *document warehouse*;
- Os *document warehouses* são concebidos por forma a permitir a integração de documentos que se relacionam com base na semântica dos seus conteúdos.

Este autor considera ainda que as cinco principais etapas a considerar aquando da criação do *document warehouse* têm correspondência às de que devem ser consideradas na implementação de um *data warehouse* (Sullivan, 2001). O paralelismo é apresentado na figura 5.

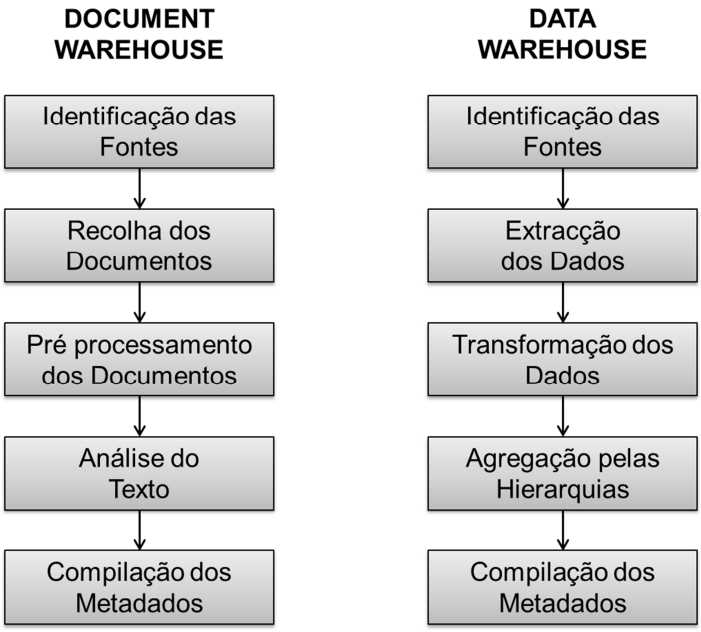

Figura 5 – Etapas na construção de *document warehouse* e *data warehouse* - adaptado de (Sullivan, 2001)

Como foi referido no capítulo inicial, o propósito do projecto visou a utilização das técnicas e processos de text mining que permitam não só minimizar os impactos das questões relacionadas com os processos de gestão de informação, mas simultaneamente, valorizar um activo como é o caso do *document warehouse* da Manchete, que inclui mais de quatro milhões de registos.

No capítulo seguinte descreve-se o ambiente no qual o projecto NovaIntell foi realizado, sendo apresentado o processo de gestão de informação implementado na Manchete, onde o *document warehouse* se inclui.

3. Projecto NovaIntell

A formulação do projecto NovaIntell foi suportada por uma estratégia de complementaridade de competências e de interesses comuns para o melhor aproveitamento dos resultados de actividades de investigação e desenvolvimento tecnológico (I&DT) mediante o estabelecimento de um contrato de consórcio entre uma empresa dedicada à gestão de informação, a Manchete, com uma entidade do Sistema Científico e Tecnológico (SCT), o ISEGI-UNL, instituição com competências reconhecidas na área de intervenção deste projecto e que neste âmbito se associaram para potenciarem sinergias, bem como partilharem custos e riscos.

O carácter de inovação associado ao projecto NovaIntell resultou da criação de novos conhecimentos e respectiva aplicação no mercado numa área emergente que tem vindo a despertar a atenção dos mais diversos sectores de actividade económica, uma vez que a exploração e a criação de valor com base em repositórios de conhecimento não estruturado são uns dos principais desafios que se colocam às organizações na sociedade da informação e do conhecimento.

3.1. Contexto Manchete

A Manchete é uma empresa que actua no sector da informação e conhecimento, tendo como missão a mitigação do grau de incerteza inerente a cada decisão estratégica, táctica ou operacional dos seus clientes. Fundada em 1996, conta actualmente com cerca de 50 profissionais efectivos, fazendo parte das associações internacionais do sector onde actua, como são o caso da Fédération Internationale des Bureauxs d'Extraits de Presse (FIBEP), The International Association for the Measurement and Evaluation of Communication (AMEC) e Strategic and Competitive Intelligence Professionals (SCIP), tendo como base de clientes empresas nacionais e internacionais de média e grande dimensão, bem

como organismos públicos. Com uma cultura centrada nas necessidades de informação e conhecimento dos seus clientes, a estratégia da Manchete tem-se focado na inovação, na qualidade e por ser pioneira nos sectores onde actua. Do seu percurso de inovação e pioneirismo destacam-se os seguintes marcos:

- 1997 – primeiro portal web no sector
- 1998 – primeira base de dados online de informação
- 1999 – primeira base de dados de media measurement online
- 2000 – primeira empresa a abolir o clipping em papel
- 2001 – canal MediaZap (TvCabo)
- 2002 – primeiro portal WAP do sector
- 2003 – adesão à FIBEP
- 2004 – adesão à SCIP
- 2006 – adesão à AMEC
- 2007 – constituição Manchete Angola
  – monitorização de web social
  – plataforma PRM | Public Relations Management
- 2008 – novo portal Mynetpress
- 2009 – primeira empresa a desenvolver Text Mining em português
  – implementação do primeiro sistema de conversão de voz para texto em português
  – parceria para a representação em Portugal da LexisNexis

Para satisfazer as diferentes necessidades dos seus clientes, mantém o seu negócio organizado nas três seguintes áreas: *Media based*, *Reputation based* e *Intelligence based*.

A actividade da unidade *Media based* assenta nas actividades de *clipping*, *measurement*, *media intelligence* e *public relations management*. A unidade *Reputation based* foca-se na análise de reputação dos conteúdos publicados. Por sua vez, a unidade *Intelligence based* dedica-se aos processos de text mining, SEW, IC, research, biografias e web social.

Apesar da Manchete não se ter dedicado ao negócio das tecnologias de informação, a empresa tem baseado toda a organização e disponibilização da informação, conhecimento e inteligência na sua plataforma tecnológica, o *Knowledge Based Portal* (KBP).

Esta plataforma tem evoluído no sentido de enriquecer a informação com vista a acrescentar valor ao negócio dos clientes. Para além de ter integradas novas tecnologias de indexação e pesquisa, administração, partilha e distribuição da informação e gestão de comunicados de imprensa, inclui conteúdos de publicações em papel, online, rádio e televisão.

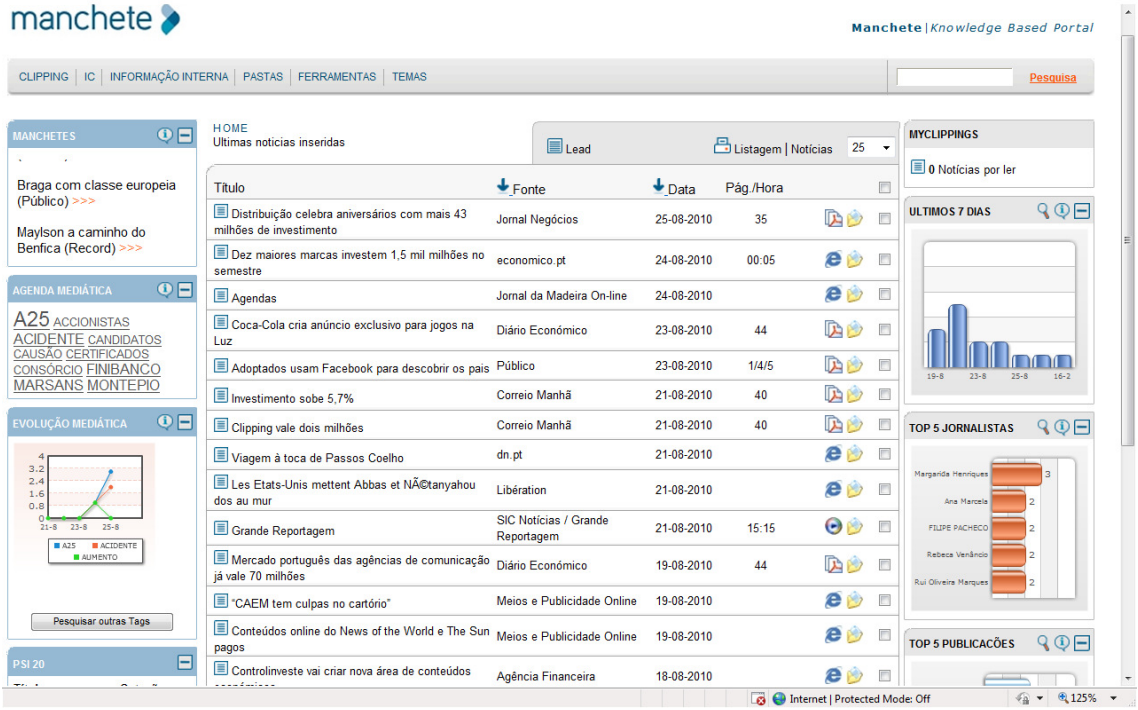

Figura 6 – Imagem do *Knowledge Based Portal* da Manchete

Nesse contexto, o KBP disponibiliza aos seus utilizadores as seguintes funcionalidades:

- Cobertura – Monitorização de imprensa escrita nacional, regional, meios online, rádio e televisão;

- Diversos – Manchetes, agenda mediática, evolução mediática e cotação dos títulos do PSI 20;
- Ferramentas
    - Acesso à Pesquisa: por temas, palavra-chave, data, etc.;
    - Acesso à Personalização: selecção de temas a aparecer na página inicial;
    - Temas – Ligação para as notícias dos últimos 30 dias relativamente aos temas seleccionados;
- Visualização da notícia em texto ou PDF e acesso ao áudio ou vídeo das notícias de rádio ou televisão;
- myClippings – Armazenamento até 30 dias das notícias para visualizar/trabalhar mais tarde;
- Ordenação – Pode-se ordenar as diferentes listagens de notícias por "Fonte" ou por "Data";
- Impressão – O centro de impressão permite imprimir automaticamente a listagem e notícias do dia, ou do dia e mais dois dias, de forma a imprimir rapidamente as notícias do fim-de-semana;
- Comentários – Funcionalidade que permite acrescentar notas, devidamente assinaladas que serão partilhadas por todos os utilizadores da empresa.
- Selecção de várias notícias, permitindo a compilação em PDF e/ou envio de notícias por correio electrónico.

O KBP é porventura, a face mais visível dos processos de tratamento de informação desenvolvidos pela Manchete. Dos seus processos de negócio fazem parte a recolha permanente de informação sobre mais de 2.000 fontes distintas entre jornais em papel, jornais em formato electrónico, rádio, televisão, imprensa nacional, internacional e regional, revistas de especialidade, entre outras. Estes conteúdos são armazenados em repositórios de dados, sendo de seguida disponibilizados a cada um dos seus destinatários de acordo com a sua temática.

Os problemas que geralmente se colocam, não apenas à Manchete mas à generalidade das empresas de gestão de informação quando tratam as notícias, são os seguintes (Peters, 2005):

- Indexação manual;
- Custo de execução;
- Tempo de execução;
- Subjectividade de quem classifica a notícia;
- Incapacidade para catalogar convenientemente todas as notícias que se pretendem arquivar.

Por dispor de uma base de dados com mais de quatro milhões de documentos em texto e por processar cerca de três mil notícias por dia, a Manchete sentiu a necessidade de mitigar os problemas enunciados e simultaneamente, explorar o potencial oferecido pelo desenvolvimento de processos de text mining com vista a gerir de forma mais eficaz o seu *document warehouse*. Foi neste contexto que a Manchete iniciou em 2001 contactos informais com a Temis, por intermédio de Alessandro Zanasi, um dos seus co-fundadores, na expectativa de desenvolver um plano empresarial focado nas potencialidades oferecidas pelos processos de text mining.

A estratégia de criação do *document warehouse* tem passado pela conversão e centralização de todos os conteúdos recolhidos numa plataforma comum – o texto em formato digital. Independentemente do formato original dos conteúdos, quer sejam rádio, televisão, imprensa escrita, os respectivos conteúdos são transcritos e armazenados numa base de dados central.

Para a conversão dos conteúdos de rádio e televisão, a Manchete dispõe de um sistema automático de transcrição dos conteúdos directamente para o *document warehouse*. Este sistema permite efectuar pequenos ajustes e correcções de algumas falhas, resultantes do processo de transcrição ser efectuado de forma totalmente automática. Neste interface, apresentado na figura 7, é possível corrigir as transcrições produzidas (transcrição, resumo e título), editar o momento de

início e fim de cada notícia, bem como alterar os temas associados a uma determinada notícia. Após efectuar as alterações é possível guardá-las numa tabela, distinta da original, preservando assim a informação original.

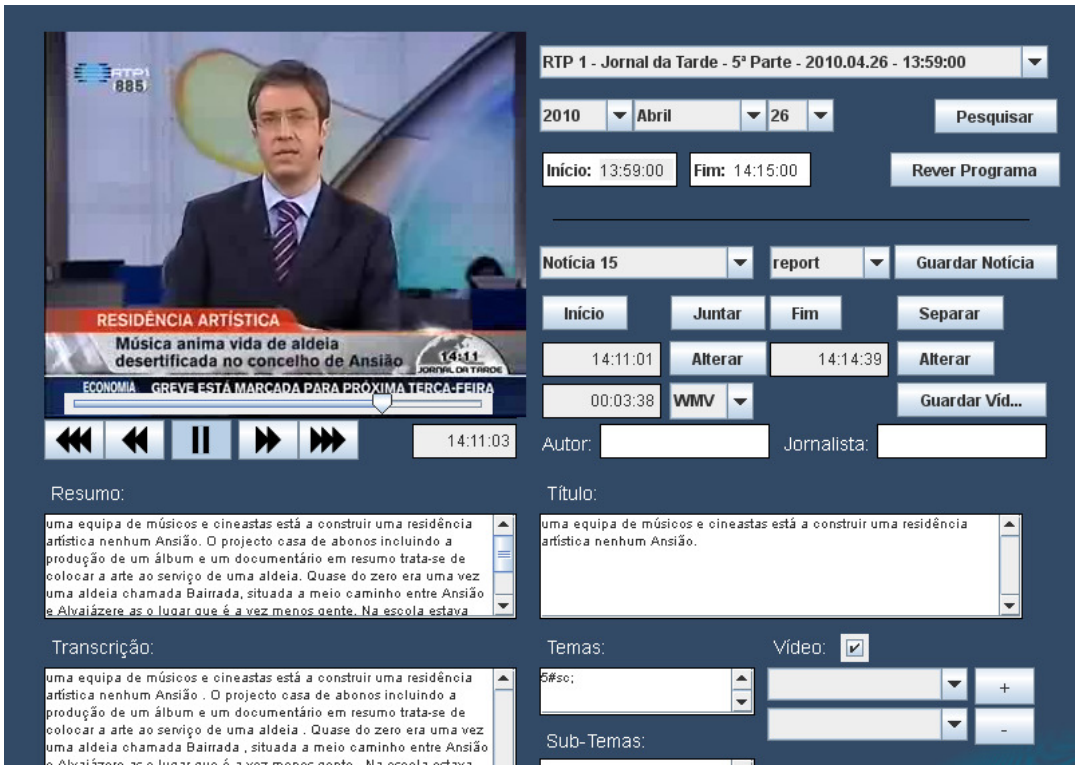

Figura 7 – Interface da aplicação de transcrição de conteúdos multimédia

Com a transcrição dos conteúdos multimédia, foi possível concretizar a uniformização da base de informação disponível para análise.

Os conteúdos com suporte em papel passam por um processo de *Optical Character Recognition* (OCR), sendo o texto revisto armazenado no repositório de dados textuais. Por seu turno, a recolha das publicações online processa-se através da utilização de aplicações que copiam os conteúdos de páginas de internet (*web crawlers*) que depositam no repositório de dados textuais os conteúdos dos meios pré-definidos, de forma automática e sistematizada.

Desta forma, a arquitectura do processo de gestão de informação implementado na Manchete contempla como fontes de dados conteúdos não estruturados tão

diversos como os provenientes de publicações em papel, jornais e revistas digitais, blogues, web social, rádio, televisão, relatórios, etc. Adicionalmente, podem ser incluídos conteúdos telefónicos como por exemplo as chamadas de um *contact center* para efeitos de análise, no âmbito de Customer Relationship Management (CRM). Como fluxo de saída do processo, para além dos conteúdos disponibilizados no KBP, estão ainda ao dispor do analista de informação conteúdos devidamente estruturados que servem de matéria-prima para a execução de análises e relatórios no âmbito de SEW, IC e AR.

O processo de gestão de informação implementado na Manchete encontra-se representado na figura 8.

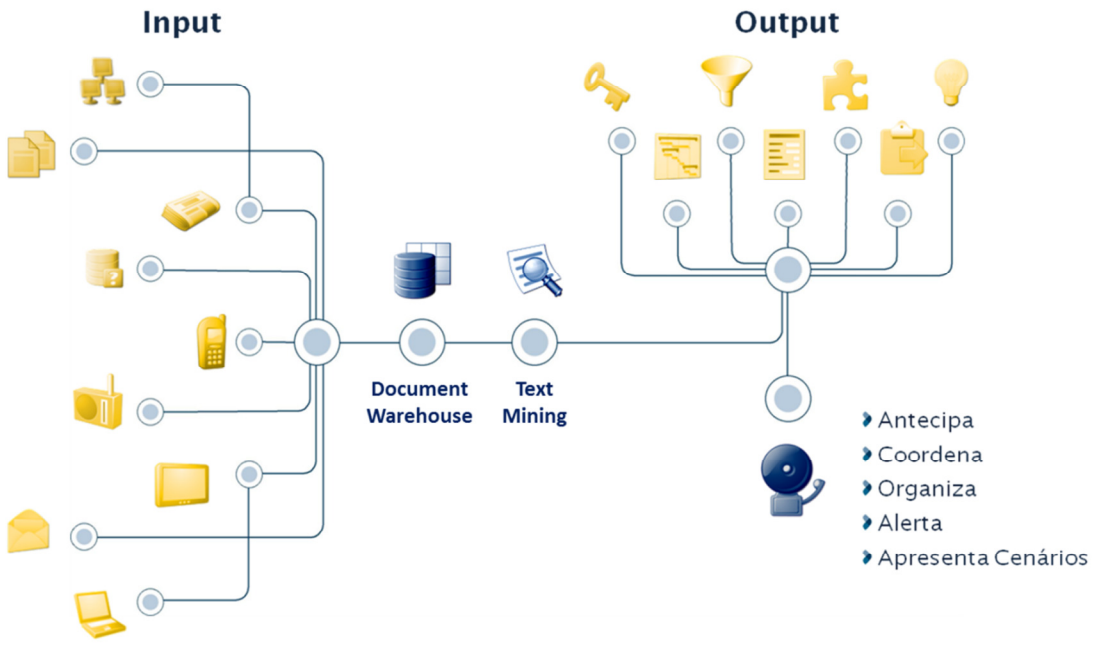

Figura 8 – Representação do processo de gestão de informação da Manchete

3.2. Introdução ao Projecto NovaIntell

A definição do plano de projecto assentou no desenvolvimento de soluções específicas para a língua portuguesa nas vertentes da estruturação da informação,

análise de relações e análise de sentimentos. Com base neste plano, a Manchete pretendeu complementar a sua oferta tradicional consolidada no *clipping*, desenvolvendo novos serviços na esfera da gestão do conhecimento capazes de alicerçar a tomada de decisão operacional, táctica e estratégica dos seus clientes.

A componente de estruturação da informação consistiu na anotação dos textos com vista à extracção de informação relevante. Este processo resultou na identificação de entidades concretas como são o caso dos nomes de pessoas, locais, empresas, números de telefone, endereços postais e electrónicos, datas, etc. De modo a potenciar a capacidade analítica baseada na extracção dos conteúdos dos textos, processaram-se os metadados associados a cada documento, embebidos nos ficheiros XML que alimentam os conteúdos no sistema. Estes metadados incluem elementos diversos tais como a fonte dos dados, data de publicação ou emissão, idioma, meio, país, data de processamento, suplemento, temática e autor. A junção das entidades com os metadados resultou numa estrutura lógica da informação textual existente nos repositórios de dados, que serve de alicerce ao trabalho necessário para a realização das análises de *intelligence* subsequentes.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <tm xmlns:dc="http://purl.org/dc/elements/1.1/" version="1.1">
  - <doc id="ea25cbd2-5706-4a76-9ec9-d9168f1ea498" href="">
      <dc:source>Diário Notícias</dc:source>
      <dc:date format="yyyy-MM-dd">2010-08-17</dc:date>
      <dc:title>Revolta no último regresso à Delphi</dc:title>
    - <text zone="news">
        <data>Revolta no último regresso à Delphi encerramento Os 321 trabalhadores da Delphi da Guarda regressaram ontem ao
          trabalho depois das férias pela última vez antes do fecho da fábrica, anunciado para o final do ano, e poucos foram os que
          conseguiram esconder a revolta e a angústia que sentem com a decisão tomada pela administração da empresa. Para
          Horário Antunes, um dos trabalhadores, não podia haver pior regresso de férias na fábrica da multinacional de cablagens
          para automóveis: "Já sabemos aquilo por que vamos passar: é quase como os animais que vão para o matadouro, mas é
          ávida", lamentava ontem, à entrada da fábrica da Guarda.</data>
      </text>
    - <features>
        <ft f="1">/Metadata/Tema/Automóvel</ft>
        <ft f="1">/Metadata/Pais/Portugal</ft>
        <ft f="1">/Metadata/Idioma/Portugues</ft>
        <ft f="1">/Metadata/Meio/Diário Notícias</ft>
        <ft f="1">/Metadata/Suplemento/</ft>
        <ft f="1">/Metadata/Fonte/Imprensa Nacional - Jornais</ft>
        <ft f="1">/Metadata/Temática/Indústria</ft>
        <ft f="1">/Metadata/Autor/</ft>
      </features>
  </doc>
```

Figura 9 – Exemplo de um ficheiro de carregamento de notícias, em formato XML

No contexto das análises levadas a cabo pela Manchete, tanto ou mais importante do que saber quais os *players* (por exemplo, pessoas e empresas) de determinado

sector de actividade é determinar a forma como estes se relacionam. Com base nas entidades identificadas no decurso do processo de estruturação da informação, a fase posterior do plano compreendeu a análise das relações que podem ocorrer entre si. Desta forma passou a ser possível não só vislumbrar a existência de relações entre entidades mas também determinar a sua tipologia. Situações típicas decorrentes da análise de relações são, por exemplo, averiguar a existência de relações entre empresas como é o caso das parcerias, aquisições ou fusões, ou entre os seus representantes como tais como relações políticas entre administradores dessas empresas, etc.

Um outro foco de análise que se pretendeu levar a cabo relaciona-se com a possibilidade de qualificar as entidades extraídas dos textos. A análise de sentimentos visa a identificação de opiniões e sensibilidade (ou favorabilidade) relativamente a marcas, produtos, organizações e pessoas. Permite uma análise de larga escala sobre conteúdos oriundos de media e web social, *contact centers*, etc. com vista à identificação de fenómenos de popularidade (positiva ou negativa) e tendências. A análise processa-se através da extracção de expressões de texto, classificando-as de acordo com a sua natureza e intensidade. Esta análise é particularmente útil em análises de marketing, particularmente em estudos de receptividade de produtos, serviços e o acompanhamento de campanhas efectuadas.

Contudo, por se tratar de um projecto extenso, a presente dissertação foca-se na etapa da estruturação da informação, o ponto fulcral do projecto por ser o alicerce das etapas posteriores. Efectivamente, a generalidade dos processos de text mining enunciados no capítulo anterior, como é o caso da categorização, só podem ser executados após ter ocorrido a devida estruturação dos documentos.

Apesar de ter havido uma propensão para utilizar as soluções da Temis em virtude dos contactos já estabelecidos e anteriormente referidos, procedeu-se a uma análise de mercado sobre as principais funcionalidades de text mining disponibilizadas pela Temis ("XeLDA White Paper"), SPSS ("Mastering New Challenges in Text Analytics", 2008) e Teragram ("European and Arabic

Linguistic Suite") para a língua portuguesa. O quadro resumo das funcionalidades apresentado na tabela 3 permite verificar que as diferenças de base entre estas não são particularmente significativas.

Temis	LexiQuest/SPSS/IBM	Teragram/SAS
Language Identification	Language Identification	Morphological Stemming
Tokenization	Document Convertion	Spelling Correction
Morphological Analysis	Segmentation (Tokenization)	Part of Speech Tagging
POS Disambiguation	POS Dictionary	Text Normalization
Text Extraction		Word and Sentence Tokenizer

Tabela 3 – Principais funcionalidades de algumas aplicações de Text Mining

Neste contexto, optou-se por escolher a solução analítica de texto cuja empresa demonstrou desde a primeira hora empenho e suporte ao desenvolvimento do módulo linguístico português. Assim, a solução de text mining adoptada pela Manchete e utilizada como ferramenta de suporte ao desenvolvimento deste projecto foi a versão 5.0 da solução Luxid, desenvolvido pela empresa francesa Temis.

Esta opção foi reforçada pelo facto da empresa Temis fazer parte da lista elaborada pela revista KMWorld das 100 empresas mais importantes no âmbito da Gestão do Conhecimento (McKellar, 2009). Mais recentemente, a sua aplicação Luxid recebeu o galardão de *Best Business Intelligence or Knowledge Management Solution* pela Software & Information Industry Association ("SIIA").

3.3. Caracterização do Projecto NovaIntell

Nesta secção são apresentados os aspectos que caracterizaram o desenvolvimento do projecto NovaIntell, tendo por base a candidatura submetida à AdI com vista à obtenção dos apoios à inovação previstos no âmbito do QREN.

Por se tratar de um projecto em co-promoção entre a Manchete e o ISEGI-UNL, o perfil profissional e académico do pessoal afecto permitiu a criação de uma equipa de projecto multifacetada, com valências desde a gestão e controlo financeiro até especialistas em data mining e BI, passando por informáticos e peritos em IC. Na repartição de responsabilidades, a equipa da Manchete teve a seu cargo a construção da solução analítica de texto, tendo a equipa do ISEGI-UNL participado no acompanhamento dos trabalhos, na co-avaliação dos desenvolvimentos executados e na promoção dos resultados alcançados.

No contexto deste trabalho de projecto é relevante explicitar as responsabilidades que o autor teve ao longo da sua execução. Essas responsabilidades foram desempenhadas a dois níveis distintos; operacional e técnica. Na óptica operacional, o autor desempenhou a coordenação operacional do projecto, tendo participação activa desde a sua formulação e formalização da candidatura aos incentivos do QREN, até à sua conclusão, com a submissão dos elementos exigidos pela AdI e elaboração do relatório de projecto. Na perspectiva técnica, o autor teve a seu cargo a construção do *skill cartridge* TM 360 para a língua portuguesa, cuja descrição é feita no capítulo 4.

Uma vez que o projecto compreendeu o cumprimento de seis etapas, apresentam-se de seguida os aspectos caracterizadores de cada uma dessas componentes.

3.3.1. Estudos Preliminares

Em virtude do arranque do projecto ter ocorrido em Julho de 2008, cinco meses após a sua elaboração e submissão da respectiva candidatura aos apoios do QREN, procedeu-se nesta fase à revisão do plano geral do projecto. Foram

analisadas e revistas algumas das suas componentes operacionais, como por exemplo, a alteração da data de realização do workshop sobre text mining que foi realizado juntamente com as restantes acções de promoção do projecto.

Pelo motivo referido anteriormente, foi reservado um período destinado à mobilização e disponibilização dos recursos humanos, técnicos e financeiros necessários para o desenvolvimento das fases posteriores do projecto.

3.3.2. Especificações Técnicas

Esta etapa, que decorreu em simultâneo com a execução dos estudos preliminares, teve como propósito a revisão das especificações do software e hardware a utilizar ao longo do período de execução do projecto. Neste contexto e tendo em conta que já havia sido efectuado o levantamento dos requisitos informáticos, esta fase serviu para comprovar a validade dos mesmos, tendo em conta uma eventual nova versão do software ou uma nova configuração de hardware, bem como para a instalação e configuração dos equipamentos e aplicativos necessários para a sua execução.

Efectivamente, no decurso do projecto, houve uma actualização da versão do software de text mining Luxid, da versão 5.0 para a versão 5.1. Apesar de se passar a dispor de funcionalidades adicionais, estas novas valências incidiram sobretudo na componente analítica, ou seja, focadas no trabalho realizado pelo utilizador final, tendo este um perfil de analista de informação. Por conseguinte, a migração do software não teve impacto no trabalho de concepção do *skill cartridge* TM 360.

3.3.3. Aquisição de novos conhecimentos

Em virtude do projecto basear-se em processos e tecnologia nova e por envolver pessoas com áreas de formação e especialização distintas foram realizadas acções de formação nas componentes fulcrais do projecto. Para além da formação sobre

as áreas de text mining e IC, foram levadas a cabo acções de formação na aplicação Luxid.

A formação em text mining, ministrada pela Dra. Anne Schwartz especialista da Temis em linguística e text mining, focou-se na aprendizagem dos aspectos relacionados com a concepção de s*kill cartridges*. Foram ainda abordadas as funcionalidades que o sistema disponibiliza ao administrador do sistema e ao utilizador numa óptica de analista de informação.

A componente de formação em IC foi ministrada pela Dra. Fátima Rebelo, por dispor de uma vasta experiência e conhecimentos inerentes à execução de projectos ao longo do seu percurso profissional e pela frequência em cursos de formação ministrados pela Academy of Competitive Intelligence ("ACI") e participação em congressos da SCIP. Os temas abordados ao longo da acção de formação em IC incluíram uma introdução à IC, o ciclo e gestão da IC e a criação de relatórios de IC.

3.3.4. Desenvolvimento

O processo de desenvolvimento foi a etapa nuclear do projecto NovaIntell. Para além de ser a etapa mais extensa, foi nesta fase que se construíram os processos de extracção e análise da informação textual escrita na língua portuguesa, com vista à utilização da informação sistematizada numa perspectiva de gestão de informação e IC.

Neste contexto, o desenvolvimento do projecto NovaIntell, assentou na execução de três macro processos com vista à criação de um modelo capaz de extrair e estruturar a informação com vista a obter conhecimento:

- Criação de dicionários de termos;
- Construção de regras linguísticas;
- Definição de modelos de IC.

A execução destes macro processos resultou na criação de um processo linguístico específico para a extracção de conhecimento sobre documentos redigidos na língua portuguesa. A apresentação detalhada do processo linguístico implementado é efectuada no próximo capítulo.

A conjugação dos dicionários de termos e regras linguísticas criadas neste processo de desenvolvimento permitiram a evolução dos modelos de IC nas seguintes perspectivas:

- Descritiva - tem como intuito aumentar a compreensão sobre um conjunto de documentos. A sua natureza é eminentemente exploratória, passando pela sintetização da informação donde resulta a identificação das entidades, a sua relação, a sumarização e agrupamento dos conteúdos presentes nos textos analisados. O objectivo é extrair informação de maneira a que se possa avaliar rapidamente os conteúdos dos textos (extraindo temas ou conceitos em textos muito extensos ou em conjuntos de documentos), assim como obter informação detalhada sobre termos, frases ou outras entidades ou procurar relações de semelhança entre documentos, agrupando-os em clusters (conjunto de documentos agregados em função de partilharem conteúdos similares ou afins) com graus de relevância diferentes e fornecendo informação sobre os conceitos subjacentes aos clusters.

  Este processo é potenciado pela possibilidade de manipular os metadados associados aos documentos processados. Elementos como a data de criação do documento, data de processamento, fonte, meio, língua, etc. complementam as valências descritivas disponibilizadas pelo sistema.

- Analítica - processo com vista à extracção de conhecimento que, numa perspectiva de IC e com base em informação anterior conhecida, pode resultar na antecipação de resultados sobre factos desconhecidos. Pressupõe a classificação dos documentos em categorias tendo em conta a confiabilidade das fontes, entre outros critérios, e a utilização da informação implícita no texto para suportar a tomada de decisões (por

exemplo, a identificação precoce das reclamações dos clientes sobre um determinado produto ou serviço, possibilitando a rápida identificação do problema e minimizando o impacto associado).

A definição de modelos de IC visou também o estudo da aplicabilidade dos modelos descritivos e analíticos referidos anteriormente a sectores de actividade donde resulte uma vantagem competitiva. Tomando como exemplo a gestão do ciclo de vida de um produto, estão criadas as bases para o desenvolvimento de processos de melhoria nas áreas de apoio ao cliente, I&D, qualidade, marketing, etc.

Particularmente do ponto de vista do marketing, estes processos podem introduzir mudanças significativas na forma como se procede à gestão de marketing. De facto, em virtude do modo como a informação é geralmente registada em inquéritos ou *focus groups* não contemplar uma área de texto livre, o que impossibilita o registo de comentários dos entrevistados em expressão em linguagem natural (McKnight, 2005), os métodos de pesquisa e estudos de mercado tradicionais podem não ser tão fiéis à realidade quanto desejado. Através da conjugação de processos de text mining e IC, os técnicos de marketing podem acompanhar de modo mais eficiente as reacções antes, durante e depois do lançamento de um determinado produto (tanto na internet, como em meios de comunicação social, através da aplicação de ferramentas de text mining a bases de dados de notícias, por exemplo).

3.3.5. Testes e Ensaios

As tarefas realizadas nesta fase do projecto visaram a validação da qualidade dos processos de text mining e IC desenvolvidos. Nessa medida, efectuou-se uma avaliação sobre as duas perspectivas distintas: técnica e de negócio.

A validação técnica esteve relacionada com qualidade dos modelos desenvolvidos sobre a ferramenta de text mining Luxid, nomeadamente o dicionário de termos e

os algoritmos de text mining. Os testes identificaram alguns erros na extracção de pessoas e respectivas funções ou profissões. Para corrigir esses erros houve a necessidade de rever o algoritmo de identificação de pessoas, nomeadamente os dicionários e regras de composição de nomes próprios, apelidos e funções.

A figura 10 apresenta um exemplo de extracção e identificação de duas pessoas após a implementação das correcções que permitiram distinguir a função da letra «e» quando parte de um nome (entre apelidos) ou na separação de duas pessoas.

Na perspectiva de negócio, a validação foi baseada na aceitação dos modelos de IC desenvolvidos e promoção com vista a uma eventual contratação dos serviços apresentados. Por conseguinte, foi realizado um estudo em colaboração com Observatório do Tráfico de Seres Humanos, organismo tutelado pela Direcção-Geral da Administração Interna sobre a problemática do tráfico de seres humanos, que serviu de ensaio aos modelos de IC. Este estudo consistiu na identificação de situações relacionadas com problemática do tráfico de seres humanos tais como a prostituição, tráfico de órgãos, tráfico de drogas, etc., tendo resultado num relatório sobre o panorama actualizado do fenómeno em Portugal.

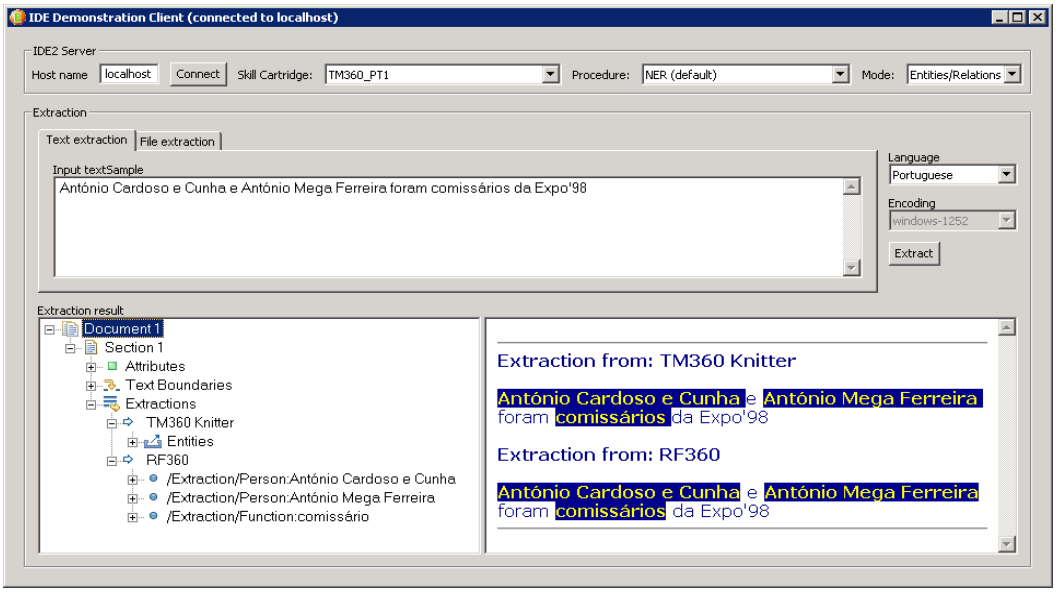

Figura 10 – Exemplo de extracção de entidades após correcção

3.3.6. Promoção e divulgação de resultados

As acções de promoção e divulgação de resultados compreenderam roadshows e workshops por diversas empresas e organizações, pela participação na 9ª Conferência da Associação Portuguesa de Sistemas de Informação ("CAPSI"), pela apresentação do projecto aos alunos do ISEGI-UNL e culminou com a realização de uma conferência na Reitoria da Universidade Nova de Lisboa, onde foram abordadas diferentes perspectivas relativas à importância da informação ao serviço da estratégia e inovação.

As entidades contactadas no decurso do périplo foram a Unidade de BI da Novabase, Collab, Direcção de Fraude da Sonaecom, Direcção Geral da Administração Interna, Unidade de Inovação, Desenvolvimento e Sustentabilidade da Galp Energia, Autoridade para as Condições de Trabalho e UMIC – Agência para a Sociedade do Conhecimento.

A participação na Conferência da Associação Portuguesa de Sistemas de Informação incluiu a redacção de um artigo científico sobre o projecto NovaIntell e uma apresentação do mesmo numa das sessões da conferência.

A apresentação do projecto à comunidade académica do ISEGI-UNL enquadrou-se na unidade curricular de Gestão do Conhecimento, em virtude dos resultados obtidos ao longo do projecto permitem novas abordagens ao estudo desta área de conhecimento.

O evento final ("Conferência Informação Estratégica e Inovação") reuniu um painel de académicos, empresários e políticos subordinada às novas formas de processar informação estratégica para as organizações teve a apresentação do projecto NovaIntell como elemento central e aglutinador. A conferência contou com um painel de oradores especialistas em text mining, inovação, energia, estratégia, linguística e análise de conteúdos.

Globalmente, a execução do projecto NovaIntell decorreu ao longo do período de tempo compreendido entre 1 Julho 2008 e 31 de Março de 2010. Até Setembro de 2009, o projecto não sofreu desvios significativos no que respeita à

calendarização. No entanto, a tarefa 5 - Testes e Ensaios, consumiu mais tempo que o inicialmente previsto devido à complexidade técnica de algumas das acções correctivas e oportunidades de melhoramento introduzidas nos algoritmos de text mining e modelos de IC. Essas correcções e validações estenderam-se até ao final de 2009, uma vez que a tarefa de promoção e divulgação apenas ocorreu em 2010.

Devido à impossibilidade de reunir todas as personalidades contactadas para a realização das sessões de promoção e divulgação dos resultados alcançados, os promotores foram autorizados pela AdI, a desenvolver estas actividades ao longo do primeiro trimestre de 2010.

A tabela seguinte compara a calendarização prevista e executada.

Tarefa	Previsto	Executado
1- Estudos preliminares	Julho 2008	Julho e Agosto 2008
2 – Especificações técnicas	Julho 2008	Julho e Agosto 2008
3 – Aquisição conhecimentos	Agosto a Outubro 2008	Agosto a Outubro 2008
4 – Desenvolvimento	Outubro 2008 a Julho 2009	Novembro 2008 a Agosto 2009
5 – Testes e ensaios	Agosto 2009 a Outubro 2009	Setembro a Dezembro 2009
6 – Promoção e divulgação	Outubro a Dezembro 2009	Janeiro a Março 2010
Projecto NovaIntell	Julho 2008 a Dezembro 2009	Julho 2008 a Março 2010

Tabela 4 – Calendarização do Projecto NovaIntell

No capítulo seguinte apresenta-se com detalhe, cada uma das fases que compõem o processo de concepção do *skill cartridge* TM 360 para a língua portuguesa, começando com a apresentação do processo linguístico construído ao longo do projecto NovaIntell.

4. Desenvolvimento do *Skill Cartridge*

Um *skill cartridge* é um módulo de software que é gerado através da codificação de um conjunto de dicionários e regras com vista à extracção de informação e conhecimento presente em textos. No caso particular do *skill cartridge* TM 360 para a língua portuguesa, este agrega um conjunto de instruções focadas no reconhecimento de entidades tão diversas como o nome de pessoas, empresas, funções desempenhadas, locais, etc.

O seu desenvolvimento foi efectuado de acordo com o processo linguístico proposto pela Temis para a criação de *skill cartridges* a utilizar na sua aplicação Luxid. Este processo linguístico é composto por duas grandes etapas; a análise morfo-sintáctica e a extracção, intervaladas por uma componente de delimitação e normalização ("Skill Cartridge Author's Guide", 2010).

As funções associadas à análise morfo-sintáctica estão assentes em dicionários, regras morfológicas e modelos estatísticos integrados na principal ferramenta de text mining da Temis – XeLDA. Esta ferramenta de transformação, normalização e extracção de informação textual foi criada pela equipa de *Advanced Technology and Systems (ATS)* do *Xerox Research Center Europe (XRCE)* com base na investigação do grupo de trabalho do *Multilingual Theory and Technology (MLTT)* detido desde Julho de 2003 pela Temis, é o motor linguístico a partir do qual são desenvolvidas as funcionalidades adicionais ("XeLDA White Paper"). Algumas das funções do processo de análise morfo-sintáctica têm correspondência com as componentes do modelo de extracção de informação abordado no capítulo 2 (Hobbs, 1993).

Por seu turno, a fase de extracção, a que corresponde a segunda componente do processo linguístico adoptado, baseia-se na construção de pequenos módulos de regras linguísticas – *Skill Units*, cada uma das quais dedicadas ao reconhecimento de entidades de um determinado âmbito, como por exemplo, o reconhecimento de locais. A agregação lógica das diversas *Skill Units* forma um *Skill Cartridge*.

A ligação entre a análise morfo-sintáctica e a extracção é estabelecida por intermédio de dois processos acessórios ao processo linguístico geral, a delimitação e da normalização dos elementos do texto. Por se tratar de elementos intermédios do processo linguístico, a sua abordagem e explicação será feita um pouco mais adiante, após o estudo da análise morfo-sintáctica.

Por conseguinte, o processo linguístico adoptado iniciado pela análise morfo-sintáctica, seguida da delimitação e normalização e concluído com a extracção pode ser representado através do diagrama da figura 11.

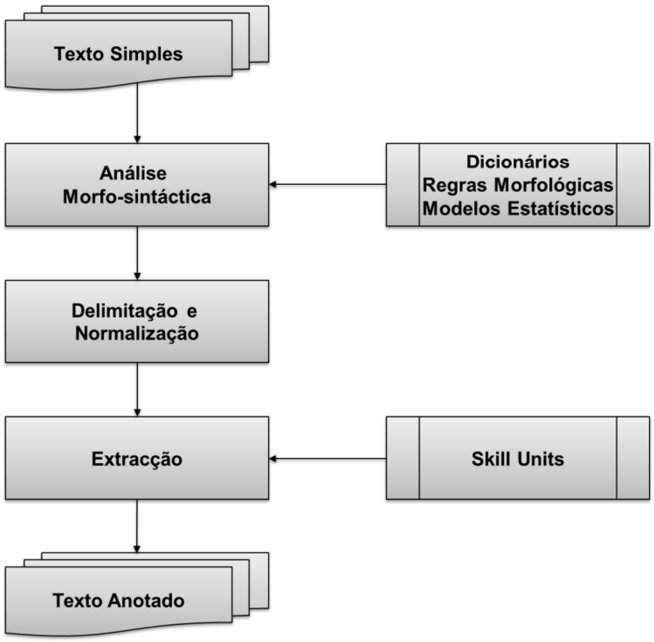

Figura 11 – Representação do processo linguístico adaptado de ("Skill Cartridge Author's Guide", 2010)

As funcionalidades linguísticas que a aplicação XeLDA disponibiliza e que alimentam a componente morfo-sintáctica do processo linguístico apresentam-se de seguida ("XeLDA White Paper").

4.1. Análise Morfo-Sintáctica

Para ser possível extrair entidades válidas é necessário efectuar-se um trabalho de tratamento prévio do texto. Este processo de tratamento inclui a identificação da língua em que os documentos se encontram escritos e separar as letras por palavras para posteriormente ser possível identificar e seleccionar a classe gramatical de cada palavra, de acordo com o contexto em que se insere. No final, são identificadas as expressões compostas por múltiplas palavras, como são o caso dos nomes das pessoas.

Da execução da análise morfo-sintáctica resultam textos estruturados, ou seja, textos cujas classes gramaticais podem ser registadas numa tabela sob a forma de linhas e colunas. As diversas componentes que compõem a análise morfo-sintáctica do processo linguístico que suporta a execução do *skill cartridge* TM 360 para a língua portuguesa podem ser representadas de acordo com a figura 12.

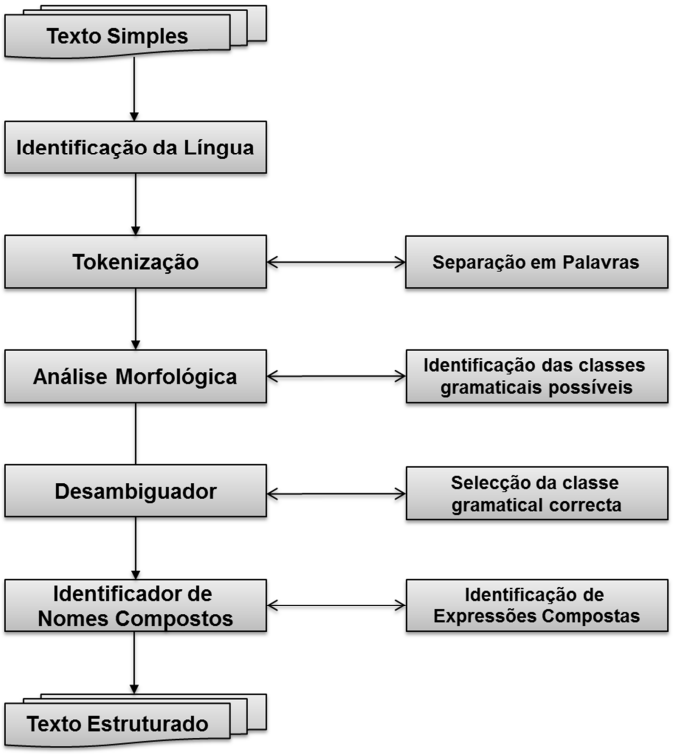

Figura 12 – Representação da Análise Morfo-Sintáctica

Descrevem-se de seguida cada uma das componentes que compõem a análise morfo-sintáctica.

4.1.1. Identificação da Língua

Esta funcionalidade permite determinar a língua em que cada documento está escrito. A sua aplicação mantém-se válida, inclusivamente, sobre textos sem acentuação. A determinação da língua em que o texto está escrito baseia-se na análise estatística dos termos mais frequentes de cada língua sobre um vasto conjunto de treino de modo a garantir a fiabilidade dos resultados.

No caso apresentado na figura 13, a expressão "*O Pedro escreveu esta frase para testar o XeLDA.*" foi classificada como sendo escrita em português com 99,94% de probabilidade. Os remanescentes 0,06% de probabilidade foram associados à designada língua espanhola, leia-se castelhano. Excluídas definitivamente como hipóteses válidas ficaram as línguas italiana, húngara, francesa, checa, holandesa, inglesa, polaca e alemã.

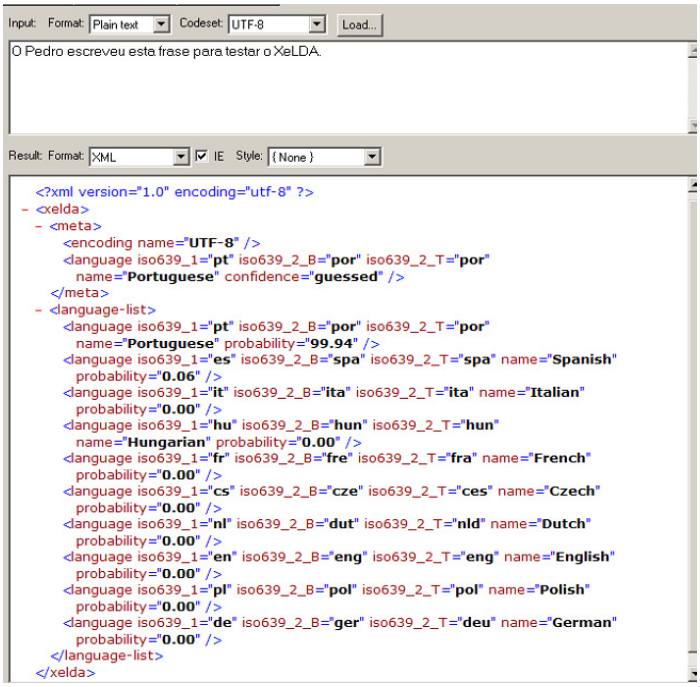

Figura 13 – Exemplo da Identificação da Língua

4.1.2. Tokenização[11]

Esta função separa o texto em palavras. Apesar de parecer uma tarefa simples que recorre a um algoritmo baseado na acentuação e espaços, tem que lidar com algumas situações particulares como das palavras ligadas por hífen (-). A palavra quarta-feira apresentada na figura 14 é um exemplo desta situação.

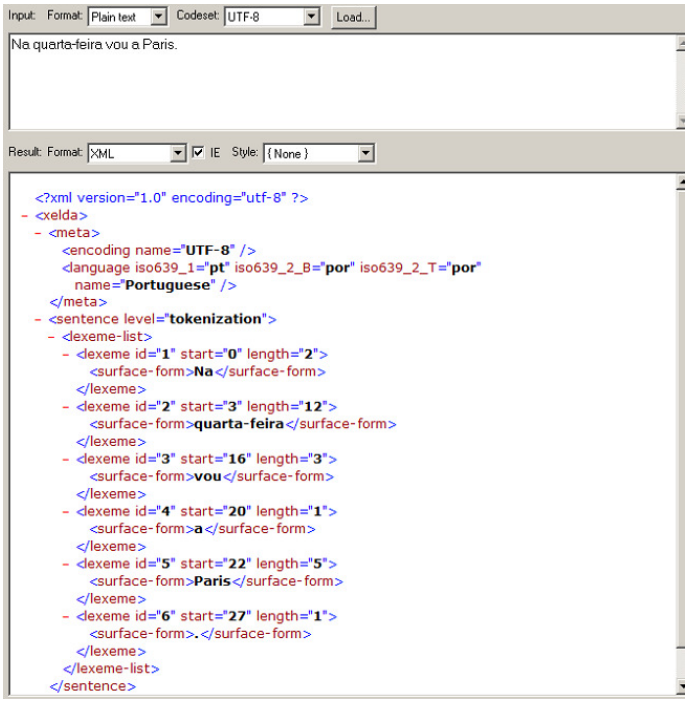

Figura 14 – Exemplo de Tokenização

Ao testar a frase "Na quarta-feira vou a Paris.", o sistema identificou cada uma das palavras, assinalando o início e a sua extensão. A palavra "quarta-feira" inicia-se na posição 3 (após a palavra "Na" e do espaço em branco), tem 12 caracteres de comprimento (6 da palavra "quarta", 1 do hífen e 5 da palavra "feira"). Foi ainda identificado o ponto final que termina a frase (início na posição 27 e tamanho unitário).

[11] Derivado do termo anglo-saxónico "token"

4.1.3. Análise Morfológica

Apresenta todas as classes gramaticais que cada palavra e respectivas derivações podem assumir no texto. No caso do exemplo da figura 15, as palavras da expressão «O Papa» podem assumir várias formas. "O" pode ser simplesmente uma letra, um pronome singular em que o "o" pode ser lido como *"ele"* ou um artigo determinante singular. Já a palavra «Papa» tanto pode corresponder a um substantivo singular, ao Sumo Pontífice neste caso, como pode tomar o valor da terceira pessoa do singular do presente do verbo papar.

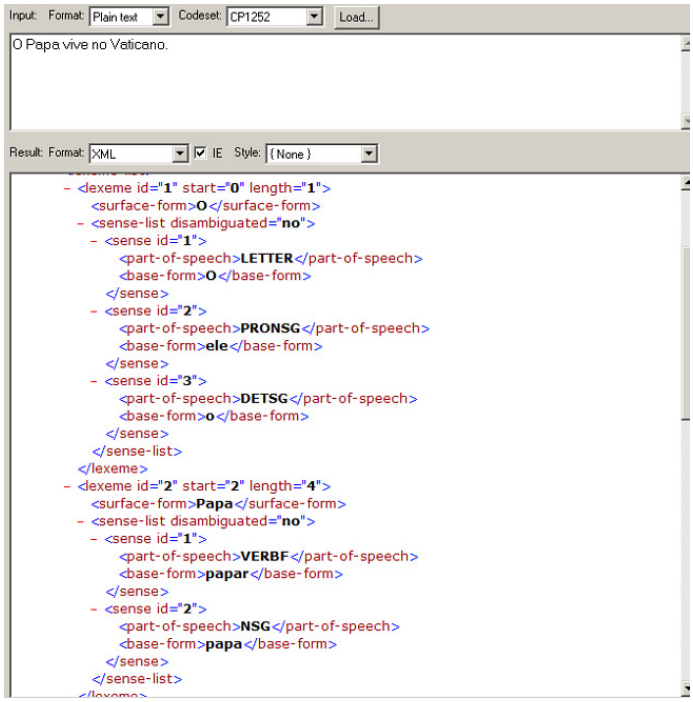

Figura 15 – Exemplo de Análise Morfológica

A selecção da forma verbal correcta em função do conteúdo e feita no processo de desambiguidade.

4.1.4. Desambiguador[12]

Esta funcionalidade foca-se na determinação da classe gramatical correcta da palavra em função do contexto em que se insere. No caso do exemplo anterior, "*o*" é considerado um artigo determinante singular e a palavra «Papa» é usada como sendo um substantivo singular e não uma forma verbal, como se constata na figura 16.

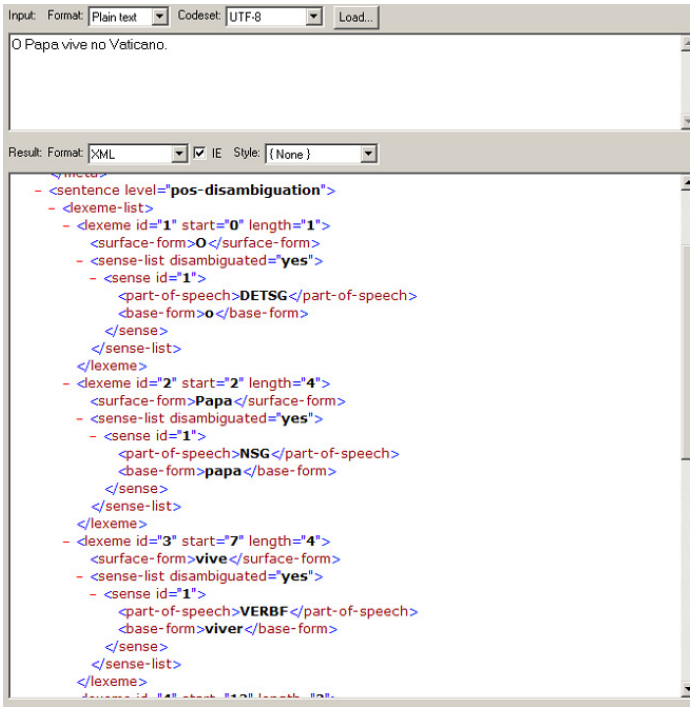

Figura 16 – Exemplo de Desambiguidade

4.1.5. Identificador de nomes compostos

Em virtude dos nomes próprios e outras entidades poderem ser compostos por mais que uma palavra, existem regras para identificar este tipo de expressões. No exemplo da figura 17, «*José Manuel Durão Barroso é o presidente da União*

[12] Derivado do termo anglo-saxónico "disambiguation"

Europeia.», são reconhecidos dois padrões resultantes de nomes compostos. O primeiro refere-se ao nome «José Manuel Durão Barroso»; o segundo padrão corresponde à «União Europeia».

As quatro palavras iniciais formam portanto o nome de uma pessoa enquanto a nona e décima palavra compõem a organização.

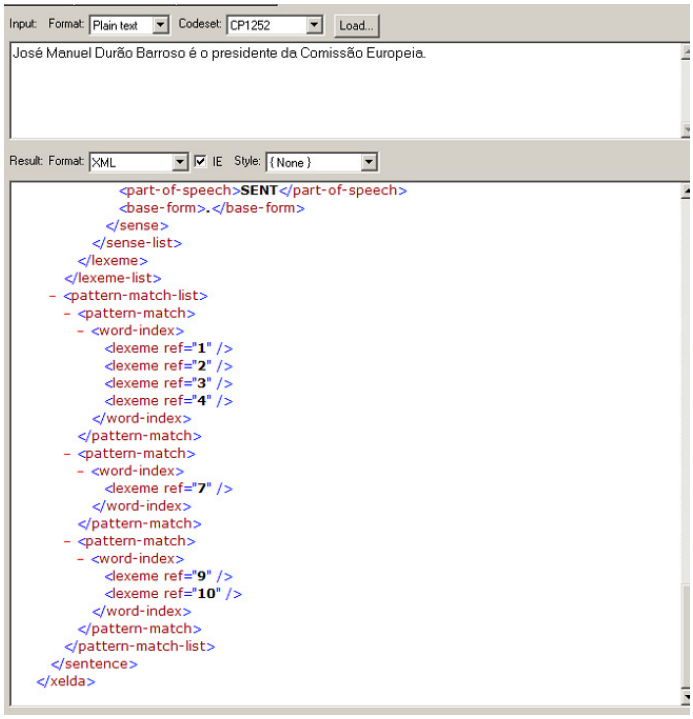

Figura 17 – Exemplo de identificação de nomes ou substantivos

4.2. Delimitação e Normalização

Como foi referido no início deste capítulo aquando da descrição do processo linguístico, existe um nível intermédio entre a análise morfo-sintáctica e a extracção. Esse nível é composto pela delimitação e pela normalização.

O processo através do qual se define o fim de uma expressão ou frase é designado por delimitação. Os marcadores de final de frase, que estão de resto pré definidos no programa, são o ponto de interrogação (?) e o ponto final (.). Apesar do ponto

de exclamação (!) não estar pré-definido no XeLDA como marcador universal de final de frase, para o caso da língua portuguesa funciona igualmente como delimitador de frase.

A normalização visa solucionar as situações das palavras ou expressões que têm múltipla grafia, como por exemplo, *cálculo* e *calculo* ou *director-geral* e *director geral*. O processo de normalização permite igualmente extrair a mesma entidade independentemente da acentuação ser usada ou não. Desta forma e em termos práticos resulta que independentemente da forma que se escreva *José Sócrates*, *José Socrates*, *Jose Sócrates* e *Jose Socrates*, todas estas expressões referem-se ao mesmo sujeito.

4.3. Extracção

No final do processo linguístico surge a extracção. É nesta última componente que se dá o reconhecimento de um determinado padrão linguístico e a respectiva classificação da entidade reconhecida. Com base na utilização das funcionalidades enunciadas nas componentes anteriores do processo linguístico e tendo em conta o grau de desenvolvimento e maturidade da aplicação XeLDA, o principal objectivo deste projecto focou-se essencialmente no desenvolvimento da componente de extracção. Em termos mais concretos, o desenvolvimento proposto foi o da criação de um *skill cartridge* que para além de permitir a estruturação de documentos, se revelasse capaz de extrair entidades a partir de conjuntos de textos escritos na língua portuguesa.

Exemplos típicos de entidades que resultam de todo o processo de reconhecimento, extracção e classificação são ("TM360 Skill Cartridge User Guide", 2007):

- Entidades universais (moradas, unidades monetárias, localizações, unidades de medida, etc.);
- Entidades do meio empresarial (empresas, organizações, pessoas, funções, etc.);

- Informações de contacto (endereços de correio electrónico, *url* de recursos disponíveis na rede, números de telefone e de fax, etc.).

O resultado visível da estruturação e extracção do texto resultante desta última componente do processo linguístico alcançou-se com a substituição dos elementos do texto, como sendo o caso das palavras, sinais e expressões, pela sua representação, ou seja, pelo nome da categoria em que essas palavras ou expressões encaixem.

Na seguinte secção apresenta-se com detalhe, cada uma das fases que compõem o processo de concepção do *skill cartridge* TM 360 para a língua portuguesa.

4.4. Criação do *Skill Cartridge*

Para que fosse possível estruturar e extrair os dados textuais foi necessário definir e codificar as regras de processamento dos textos. Em função da quantidade, diversidade e complexidade das regras inerentes a todo este processo de criação do *skill cartridge*, estas foram agregadas em parcelas definidas em função do tipo de informação a extrair. O *skill cartridge* é portanto, o agregado dessas unidades de regras e códigos mais particulares, as *skill units*.

Por sua vez, cada *skill unit* é composto pelos seguintes tipos de elementos ("Skill Cartridge Author's Guide," 2010):

- Dicionários ou listas de termos que contextualizam e atribuem significado às palavras ou expressões contidas no texto;
- Regras de extracção que definem o modo como os dicionários, listas de termos e outras componentes de código se relacionam entre si e como estas deverão ser aplicadas.

Neste contexto, um *skill cartridge* e respectivas componentes, as *skill units*, podem ser representados de acordo com o esquema da figura 18:

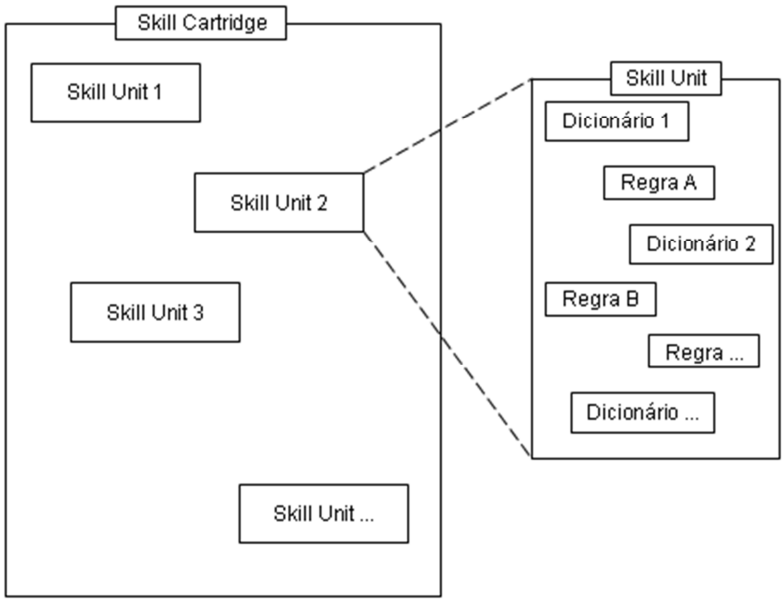

Figura 18 – Representação do *Skill Cartridge*

4.4.1. Skill Units nucleares

O ponto de partida para a criação do *skill cartridge* TM 360 para a língua portuguesa foi a criação das *skill units* com a informação linguística primária. Dessas unidades, há que destacar a *skill unit* que associa as *Part Of Speech (POS) Tagset* (ver Anexo I) em classes gramaticais. Desta associação resultou a hierarquização entre as classes gramaticais e as respectivas sub-classes. Exemplo dessa hierarquização é apresentado na figura 19, onde a agregação dos nomes comuns, nomes singulares e nomes plurais na subcategoria de *Nomes Comuns*. Esta subcategoria associada à dos *Nomes Próprios*, composta apenas por este subtipo de substantivos, deu origem ao conjunto dos *Nomes*.

```
<macro name="NOUN" display="no">
  <macro name="CommonNoun" display="no">
    <e>
     #NOUN     <!-- The same in Spanish and English.  -->
    | #NPL     <!-- Similar to NOUNPL in Spanish.  -->
    | #NSG     <!-- Similar to NOUNSG in Spanish.  -->
    </e>
  </macro>
<macro name="ProperNOUN" display="no">
   <e>
    #NPROP     <!-- Similar to PROP in English.  -->
   </e>
</macro>
</macro>
```

Figura 19 – Excerto da *Skill Unit* onde se definem as classes gramaticais

Executando o processo para as restantes classes gramaticais, como são o caso dos pronomes ou verbos, obteve-se uma *skill unit* onde foram definidas todas as classes gramaticais a ser usadas na concepção das demais *skill units* do *skill cartridge* para a língua portuguesa.

Uma outra *skill unit* crucial para o processo de criação do s*kill cartridge* para a língua portuguesa foi aquela que definiu os principais termos ou expressões usados para explicitar um determinado conceito. A título de exemplo, considere-se o caso das expressões de concordância (expressão correspondente ao inglês "according to") apresentadas na figura 20.

```
<macro name="PREP_ACCORDING_TO" display="no">
  <e>

 segundo / (a|o|as|os)
 | de / acordo / com
 | em / concordância / com

  </e>
</macro>
```

Figura 20 – Excerto da *Skill Unit* onde se definem as principais expressões linguísticas

Para além das expressões de concordância esta *skill unit* foi complementada por outras unidades de código, *macros*, com o intuito de caracterizar os seguintes conceitos derivados das expressões anglo-saxónicas:

- Sobre – derivado de *about*
- Sob – derivado de *under*
- Como – derivado de *as*
- Em – derivado de *at*
- Por – derivado de *by*
- Em – derivado de *of*
- Em ou No – derivado de *on*
- De – derivado de *from*
- Para - derivado de *to*
- Com - derivado de *with*
- Em ou Na – derivado de *in*
- Para – derivado de *for*
- Entre – derivado de *between*
- Contra – derivado de *against*

Com a construção das restantes *macros* necessárias para a arquitectura das principais expressões linguísticas, ficaram criadas as condições efectivas para o desenvolvimento do *skill cartridge*. Esta situação resulta da possibilidade das *macros* poderem ser invocadas em quaisquer outras *skill units*, no pressuposto de serem cumpridas as regras e procedimentos de invocação.

4.4.2. Skill Units regulares

Após a criação destas *skill units* nucleares, iniciou-se a concepção de unidades de código baseadas em regras, dicionários e listas de termos e expressões com o intuito de identificar e extrair entidades das mais diversas áreas do conhecimento. No entanto, por se pretender criar um *skill cartridge* que fosse simultaneamente genérico e abrangente e que por outro lado fosse capaz de processar informação

difundida pelos meios de comunicação social portugueses, foram desenvolvidas *skill units* focadas no reconhecimento e extracção de informação relacionada com as seguintes áreas:

- Empresas e organizações: definição de regras para o reconhecimento de empresas e criação de listas de empresas de referência de vários sectores de actividade;
- Contactos: definição de regras para a captação dos endereços postais, dos contactos telefónicos e de fax, dos endereços de correio electrónico assim como dos *url* dos sites corporativos;
- Eventos: criação de regras para identificar eventos, congressos, seminários, fóruns e afins;
- Localização Geográfica: dicionários dos continentes, países, províncias ou distritos e principais cidades do mundo, para além de dos municípios e localidades portuguesas;
- Media: dicionários com os principais órgãos de comunicação social;
- Pessoas: dicionários dos principais nomes próprios e apelidos e criação de regras relativas à composição dos nomes, salvaguardando a utilização de títulos académicos e honoríficos juntamente com os nomes;
- Índices bolsistas: dicionários com os principais mercados de capitais e respectivos índices;
- Unidades de Medida: definição de regras para identificação das grandezas referentes às principais unidades de medida;
- Dinheiro: definição de regras para identificação dos valores apresentados nas principais divisas;
- Números: regras para a identificação de números, quantidades, etc.

Num outro contexto de análise ou de conteúdos a processar pode surgir a necessidade de criar *skill units* complementares, capazes de recolher informação relativas a áreas de interesse específicas. Esta necessidade ocorreu na segunda fase do trabalho realizado no âmbito dos fenómenos associados à problemática do

tráfico de seres humanos, em virtude do objectivo da análise ser muito particular e dos conteúdos textuais usados serem anúncios publicados nos órgãos de comunicação social escrita, de cariz sexual com uma terminologia muito própria.

Apesar de não ser viável reproduzir os dicionários de termos e regras linguísticas construídos, apresentam-se de seguida exemplos das regras utilizadas na codificação do *skill cartridge.*

No primeiro exemplo (figura 21) apresenta-se a imagem de um excerto de um dicionário utilizado na identificação de funções políticas. Ressalva-se o facto de apesar de ser um dicionário de termos referentes a funções políticas, a metodologia usada aplica-se à generalidade dos dicionários criados.

```
<concept name="FunctionPolitical_noun" display="never" level="2">
 <e>
 ministro | embaixadora? | embaixatriz
 | deputado / (~~PREP_OF / ~~legislative-lex)?
 | presidente / ~~PREP_OF / republica
 | senadora? / (~~PREP_OF / ~~legislative-lex)?
 | primeiro / ministro
 | presidente / (~~PREP)? / ~~legislative-lex
 | secretário / geral
 | chanceler / (~~PREP / Estado)?
 | :vice/ presidente
 | mayor
 | Chefe / ~~PREP / (Estado | ~~legislative-lex)
 | (secretátio|ministro) / ~~PREP / Estado
 | diplomate#NOUN_(SG|PL)
 | (:o | :a) / numero / ~~Number$[1,2] / :do / (:partido | ~~legislative-lex)

 | presidente /~~PREP_OF / câmara
 | presidente / ~~PREP_OF / junta / ~~PREP_OF / freguesia
 | presidente / ~~PREP_OF / assembleia / ~~PREP_OF / freguesia
 | presidente / ~~PREP_OF / assembleia / municipal
 | vereador
 | presidente / ~~PREP_OF / governo / regional / ~~PREP_OF / (Região / Autónoma / ~~PREP_OF)? / (Madeira|Açores)
 | governador / civil
```

Figura 21 – Excerto de um dicionário de termos referente a funções políticas

Pegando no caso da instrução "deputado / (~~PREP_OF) / ~~Legislative-lex)?", verifica-se são chamadas as *macros* "PREP_OF" e "Legislative-lex" que são outros dicionários de termos. A *macro* "PREP_OF", que faz parte das *skill units* nucleares, é composta por diversas preposições como são o caso de: de, da, dos, das. A *macro* "Legislative-lex" faz parte de uma *skill units* regular e consiste num dicionário de termos relativos a organizações de cariz político, donde fazem parte as expressões "Assembleia de República" e "Assembleia Legislativa", entre

outras. A utilização do ponto de interrogação possibilita que a expressão entre parêntesis que o antecede seja opcional. Nestas condições, exemplo de expressões que podem ser extraídas do texto são as seguintes:

- Deputado;
- Deputado da Assembleia da República;
- Deputado da Assembleia Legislativa.

Seguindo o exemplo anterior e com base no referido dicionário, criaram-se regras para a identificação das funções políticas, como as que se apresentam na figura 22.

```
<concept name="FunctionPolitical_lex" display="never" case="preserve" level="3">
    <e>
    ~~FunctionPolitical_noun / (~~PREP_OF)? / (~~ministries-lex|~~LocationCore_NOUN)?
    | (presidente|vice/presidente|portavoz|porta/voz|vice/presidente) / ~~PREP_OF / ~~LocationCore_NOUN
    </e>
</concept>
```

Figura 22 – Exemplo de regras para a extracção de funções políticas

Através da aplicação das regras ao dicionário de termos, o sistema passou a ser capaz de identificar expressões como “deputado da Assembleia Legislativa da Madeira” ou “presidente de França” que surjam num texto e assinalá-las como funções políticas.

Já na situação relativa à identificação de pessoas, para além de serem criados dicionários com nomes próprios e apelidos mais comuns na sociedade portuguesa, a identificação de pessoas baseada apenas no cruzamento dos nomes constantes no dicionário seria redutora e ineficiente. Para resolver este constrangimento, foram criadas regras linguísticas que permitiram identificar pessoas cujos nomes próprios ou apelidos não constem nos respectivos dicionários. Essas regras assentaram na composição do nome das pessoas e na análise do contexto em que os nomes das potenciais pessoas surgem.

No caso da composição dos nomes, foi necessário contemplar as diversas formas em que o nome de uma pessoa pode ocorrer num texto. As regras iniciais foram

desenvolvidas no intuito de identificar as situações mais comuns como são o caso da ocorrência do nome próprio e apelido, dois nomes próprios e um apelido, um nome próprio e dois apelidos, apelidos separados por «e» e «de» (como o descrito no exemplo da figura 10), entre outras. Foi ainda necessário levar em consideração as situações de palavras, como por exemplo "Duarte", poderem ocorrer tanto como nome próprio ou como apelido. A figura 23 apresenta excerto da codificação utilizada na *skill unit* referente à identificação de pessoas através da composição de nomes.

```
<concept name="firstPerson" display="never" level="1">
    <concept name="Person-patt">
        <concept name="Person-lev1">
        <e>
<!-- Original:--> ~~TitleSequence / (FirstName: (~~PotentialProperN)+) / [VIX]+  <!--    ~~TitleSequence / (D\.?|Dom|Dona)? / (FirstName: (~~Potentia
<!-- Acrescentado       | ~~TitleSequence / (D\.?|Dom|Dona)? / (FirstName:(~~PotentialProperN-Min3Letters|~~PotentialProperN|~CommonAmbigFirstName|~~
<!-- Acrescentado       | ~~TitleSequence / ((FirstName:~~FirstName))$[1,2]  / [VIX]*  -->
    | (Title: (~~AmbigTitle)+) / (FirstName: (~~StrongPotentialPersonName)+)/ [VIX]+
    | (FirstName: (~~FirstName)+)  / (FirstName:(~~Initial)*) / ((FirstName:~~StrongInitial))+ / (FirstName: (~~FirstName)+) / [VIX]+
<!--    | (FirstName: (~~FirstName)+) / ((~~PREP_OF|~~Conj)? / (FamilyName: ~~PotentialProperN))+    -->
    | (FirstName: (~~FirstName)+) / (FamilyName: (~~PREP_OF|~~Conj)? / (~~FamilyName|~~LastName|~~PotentialProperN)+)
<!-- Acrescentado -->   <!--    | (FirstName: (~~FirstName))+ / (FamilyName: (~~PARTICULE~~5lprep|~~Conj)? / (~~Last_Name|~~PotentialProperN))+
<!-- Acrescentado -->   <!--    | (FamilyName: (~~Last_Name)+ / ((~~PARTICULE~~5lprep|~~Conj) / (~~Last_Name|~~FirstName)+))     -->

        |(~~TitleSequence)? / ~~KnownPersonName
        </e>
        </concept>
    </concept>
</concept>

<concept name="firstPerson" display="never" level="2" case="preserveFirst">
    <concept name="Person-patt">
        <concept name="Person-lev2">
        <e>
        <!-- Jean-Paul II , Mohamed IV   -->
         (~~TitleSequence)? / (FirstName: (~~FirstName)+ ) / (~~FamilyName| (FamilyName: ~~CompanyName-lex ) )? / ([VIX]+ | 1er)?
<!-- Acrescentado   | (~~TitleSequence)? / (FamilyName: (~~FamilyName+ / (~~PREP_OF|~~Conj)? / ~~FamilyName+)) / [VIX]* -->
        </e>
        </concept>
    </concept>
</concept>
```

Figura 23 – Excerto da *skill unit* utilizada na identificação de pessoas através da composição de nomes

A utilização de dicionário de termos e regras de composição de nomes não garantem por si só a extracção de todos os nomes de pessoas que surge nos documentos. Nessa medida foram utilizadas regras que analisam o contexto em que determinadas palavras são redigidas nos documentos. A utilização do contexto aplica-se na identificação de palavras que não constam de dicionários de nomes ou que não cumpram as regras de composição de nomes criadas mas que possam ser nomes de pessoas. A título de exemplo, se uma palavra ou sequência de palavras iniciadas por letra maiúscula surgir antes ou após um verbo, é

provável que seja o nome de uma pessoa. Neste caso, se o texto tiver a expressão "… Zinedine Zidane disse…", Zinedine Zidane será uma pessoa.

A figura seguinte mostra um excerto de uma linha de código onde se contextualiza a extracção de uma pessoa com base no contexto. Neste caso, se antes do potencial nome ocorrer um verbo, adjectivo, determinante, advérbio, etc. e se for cumprida uma determinada regra de composição do nome, então o sistema assume que aquelas palavras referem-se ao nome de uma pessoa.

```
<!-- J. Dupont -->
|  (~~DET | ~~VERB | ~~ADJ | ~~AUX | ~~ADV | ~~PREP | ~~DELIMIT | #BOA | ~~COORD_LEX | #NEG| ~~lowercaseCommonNoun) /  ( (FirstName:(~~Initial))
```

Figura 24 – Excerto de regra para identificação de pessoas com base no contexto

É ainda de referir que as regras de composição e contexto implementadas foram definidas com parcimónia, por forma a prevenir a extracção abusiva de falsas entidades. Ou seja, não podem ser criadas regras linguísticas demasiado genéricas ou abrangentes por originarem a identificação de falsas entidades.

Os exemplos anteriores retratam de modo simples como as *skill units* foram sendo criadas. A aplicação desta metodologia às restantes áreas de conhecimento referidas anteriormente originou a criação das restantes *skill units*. A figura 25 apresenta um excerto da árvore das *skill units* que foram desenvolvidas ao longo do projecto NovaIntell. A agregação das *skill units* consubstanciou-se no *skill cartridge* TM 360.

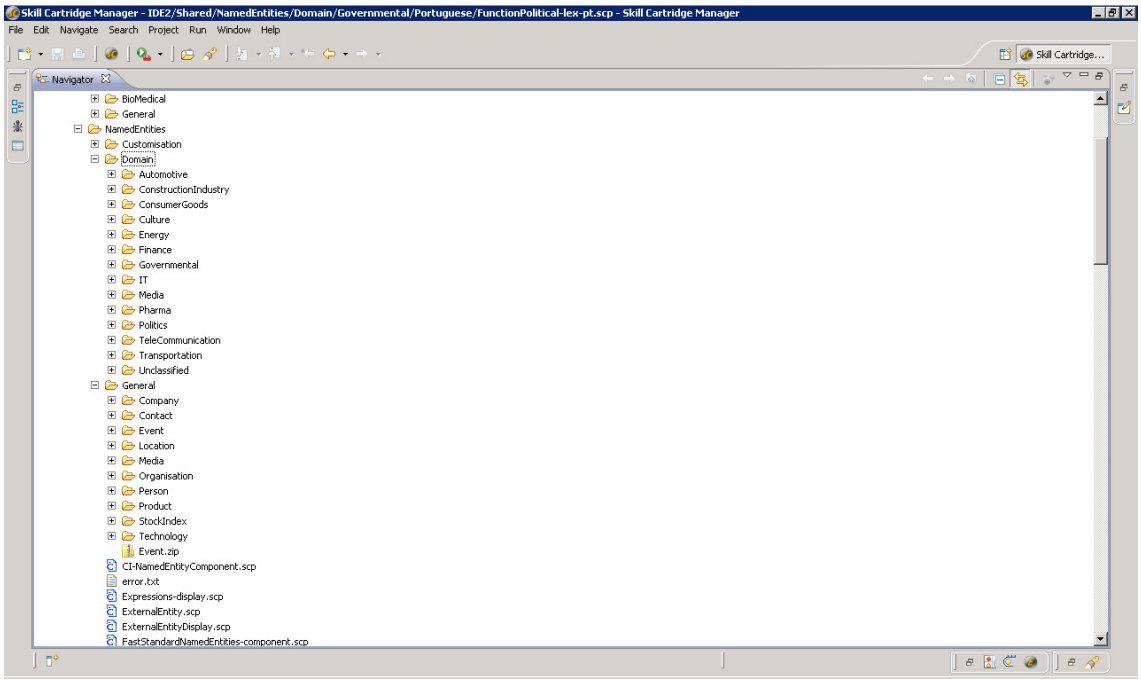

Figura 25 – Árvore das *Skill Units*

A implementação de todas estas estruturas fizeram deste *skill cartridge* para a língua portuguesa uma ampla base de trabalho através da qual se conseguiu extrair valor de documentos genéricos, como são o caso das notícias difundidas pela generalidade dos órgãos de comunicação social. Nestas condições, o presente *skill cartridge* pode servir de base para a definição de outros *skill cartridges* focados na estruturação e extracção de conhecimento sobre documentos específicos de uma qualquer área do conhecimento ou de actividade económica. Por exemplo, se for pretendido extrair conhecimento de bases de dados de registos médicos, pode-se adaptar o *skill cartridge* TM 360 de modo a acomodar dicionários de especialidades médicas, de medicamentos, etc. Do mesmo modo, podem ser criadas regras para a extracção de moléculas ou elementos activos de medicamentos.

No capítulo seguinte apresentam-se os resultados da aplicação do *skill cartridge* português a um conjunto de notícias publicadas pelos órgãos de comunicação social portugueses cujo conteúdo se encontra no repositório da Manchete.

Pretende-se assim demonstrar a capacidade do *skill cartridge* na identificação, extracção e apresentação de conhecimento baseado em textos escritos na língua portuguesa.

5. Resultados obtidos

A aplicação do *skill cartridge* TM 360 para a língua portuguesa ao software Luxid da Temis resultou numa ferramenta de visualização, análise e extracção de conhecimento. O utilizador desta solução pode pesquisar, filtrar, analisar e partilhar a informação recolhida, bem como utilizá-la para a produção de estudos na óptica da IC.

Partindo do conjunto das funcionalidades ao dispor do utilizador, tornou-se pertinente referir aquelas que poderão ser da maior utilidade para o utilizador final, ou seja, para o analista de informação. Neste contexto, destacam-se as pesquisas rápidas e avançadas, a identificação e destaque das entidades extraídas no conteúdo do texto, a análise pelas várias dimensões de entidades, a análise de proximidade das entidades, o explorador de conhecimento, a funcionalidade de partilha de informação e a actualização de relatórios ("Luxid 5.0 User Guide", 2008). Contudo, uma vez que a análise das funcionalidades da aplicação Luxid excede o âmbito do projecto, estas não serão alvo de uma abordagem exaustiva.

Como foi referido, o foco do projecto assentou na identificação das entidades presentes num conjunto de textos e as relações que estas estabelecem entre si conforme o exemplo seguinte. Após ter sido sujeito ao processo linguístico descrito no capítulo anterior, a informação extraída encontra-se devidamente identificada e assinalada, como se pode constatar no exemplo apresentado na figura 26.

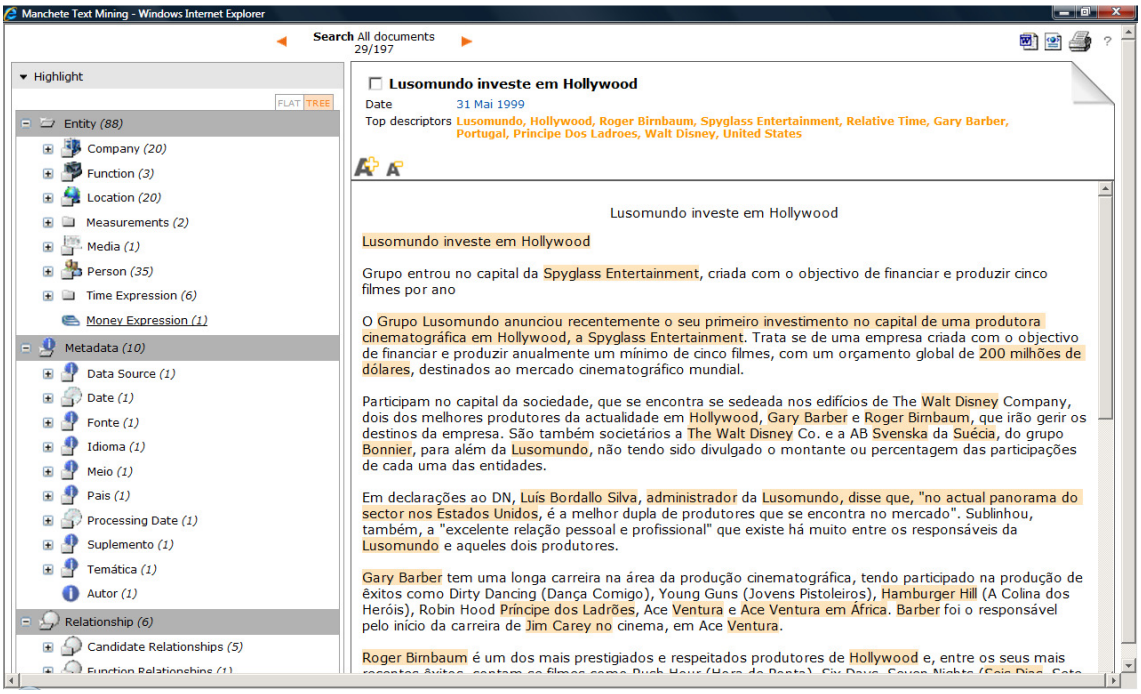

Figura 26 – Exemplo de um texto anotado e respectivos metadados

No exemplo apresentado é possível identificar entidades como empresas (Lusomundo, Spyglass Entertainment), funções (administrador), localizações (Hollywood), pessoas (Luís Bordallo Silva, Roger Bimbaum) ou valores monetários (200 milhões de dólares). É ainda possível consultar os metadados associados à notícia (fonte, meio, data, língua, temática, autor) e vislumbrar relações entre entidades.

Neste exemplo é ainda possível identificar uma relação de investimento entre a empresa Lusomundo e Hollywood. A prova dessa relação encontra-se na primeira frase da notícia "Lusomundo investe em Hollywood".

5.1. TM 360 no Luxid

Como forma de demonstrar os resultados decorrentes do desenvolvimento do *skill cartridge*, foram carregadas no sistema 14.875 notícias processadas pela Manchete nos primeiros quatro meses de 2010.

As entidades extraídas dos textos foram organizadas por categorias, em concordância com as áreas de conhecimento apresentadas no capítulo anterior.

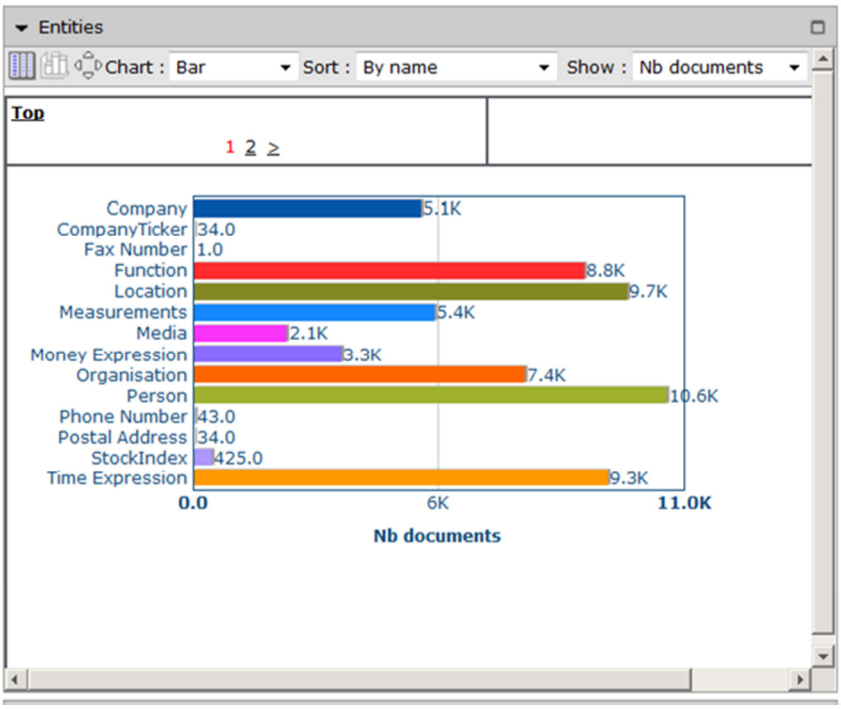

Figura 27 – Categorias das entidades extraídas

Com base na figura 27, foi possível verificar que o sistema identificou cerca de 5.100 entidades como tendo o perfil de serem empresas, 8.800 funções ou profissões, 9.700 localizações, 10.600 nomes de pessoas, 7.400 organizações, entre outras entidades. Com vista a enfatizar as capacidades de estruturação, apresentam-se de seguida exemplos de empresas extraídas pela aplicação.

A figura 28 apresenta alguns bancos extraídos do conjunto de notícias processadas. Verifica-se que "Banco de Cabo Verde" surge em quatro documentos, "Banco de Espanha" surge em dois documentos e "Banco de células do cordão umbilical" surge em apenas um documento.

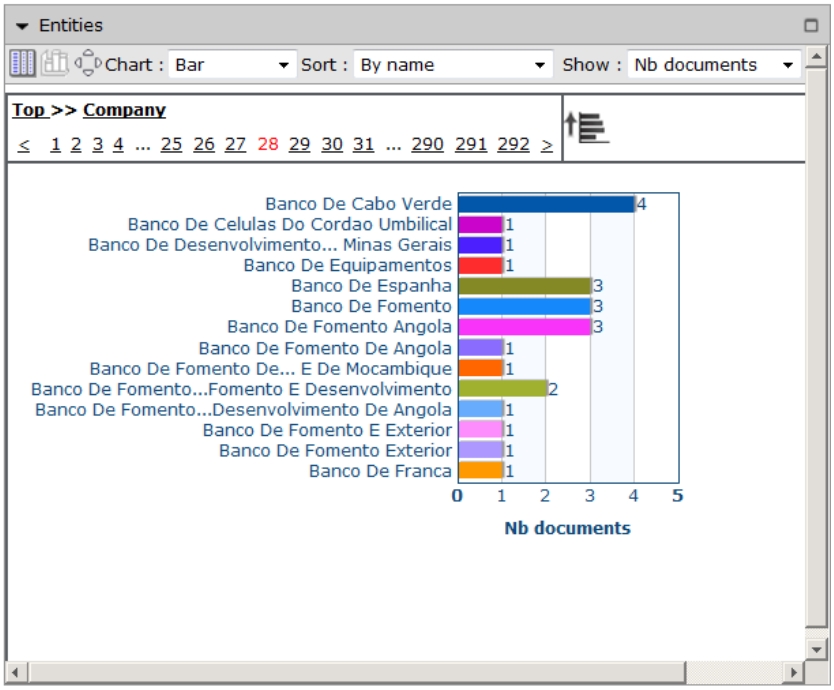

Figura 28 – Exemplo das empresas extraídas

Um outro exemplo de informação útil e valiosa para o analista de informação prende-se com a identificação e extracção das funções desempenhadas pelos indivíduos, como as apresentadas na figura 29. Neste caso, foram identificados diversos tipos de professores (economia, economia e gestão, matemática, relações internacionais, etc.).

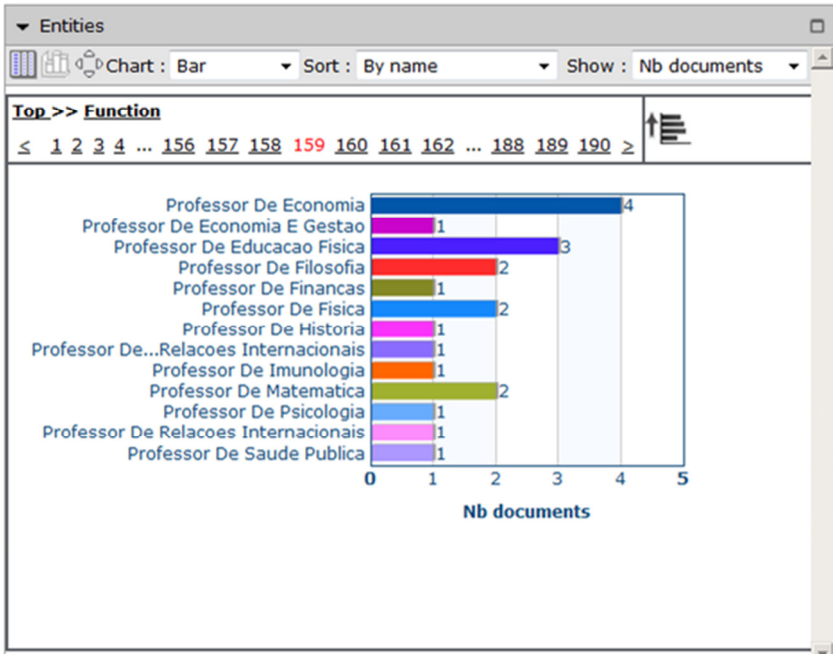

Figura 29 – Exemplo das funções identificadas

Outro tipo de informação especialmente relevante no contexto da análise resultou da identificação de pessoas. A capacidade de identificação alcançada fica patente através do exemplo apresentado na figura 30, onde se destaca a extracção de nomes pouco frequentes e estrangeiros como são o caso de Zeferino Augusto, Zeinal Bava, Zélia Duncan, Zinedine Zidane, Zinho Antunes ou Zhu Xiao Mei.

Note-se que a extracção destas pessoas com nomes pouco triviais na cultura portuguesa só foi possível graças à utilização das regras de composição de nomes e de análise de contexto referidas no capítulo anterior.

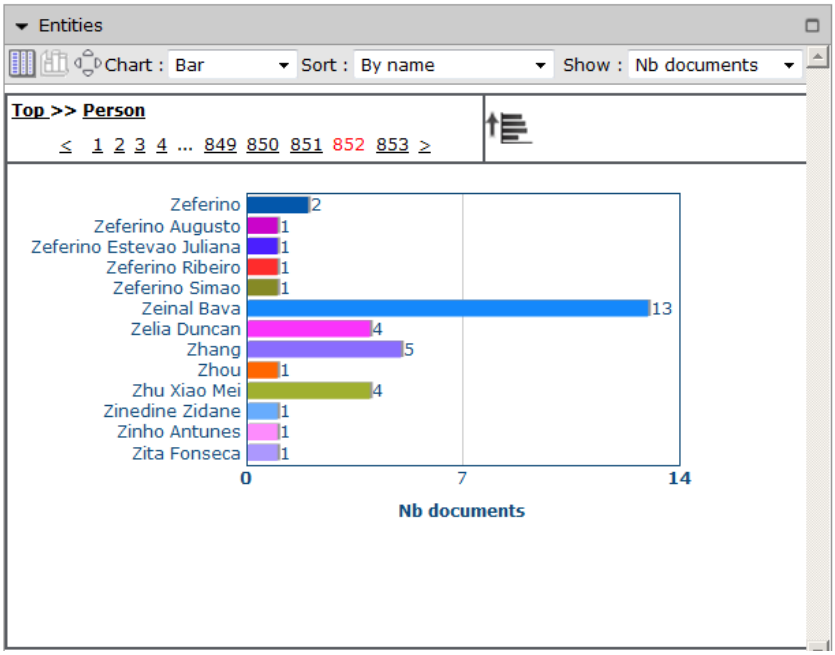

Figura 30 – Exemplo das pessoas identificadas

Com base nesta estruturação e extracção da informação, o programa foi capaz de disponibilizar a representação gráfica das entidades e a relação de proximidade existente entre si. Um exemplo da representação geral é apresentado na figura 31.

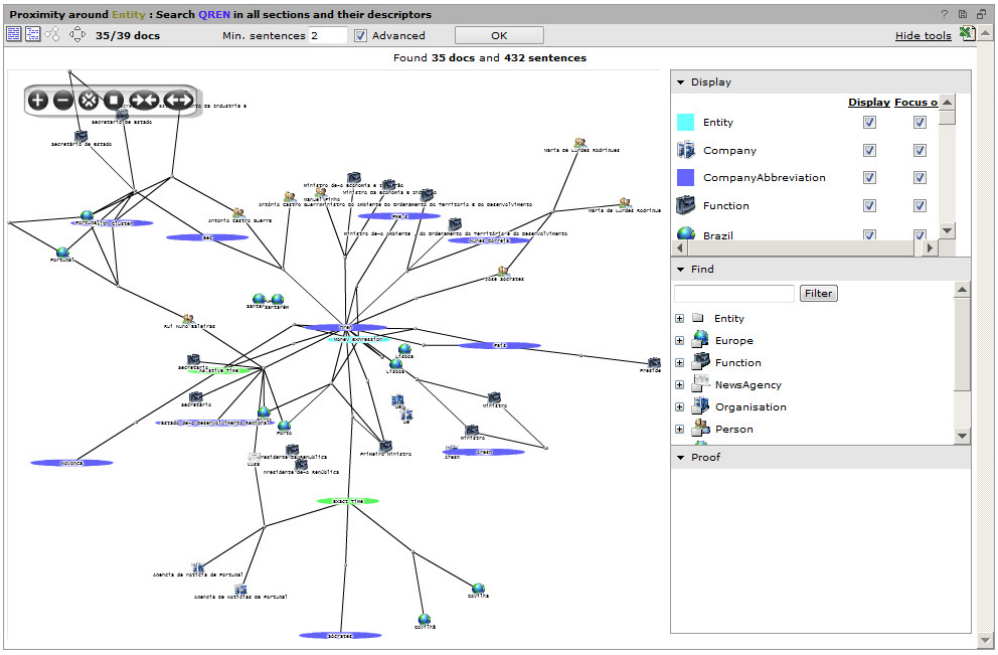

Figura 31 – Perspectiva da análise de proximidade

Para além da panorâmica geral, o sistema permite uma análise mais fina da proximidade entre entidades, apresentado a porção de texto que prova a relação de proximidade entre as entidades extraídas. Neste exemplo, verifica-se a existência de uma relação de proximidade entre o indivíduo, Rui Nuno Baleiras, e o cargo de Secretário de Estado, função por si desempenhada.

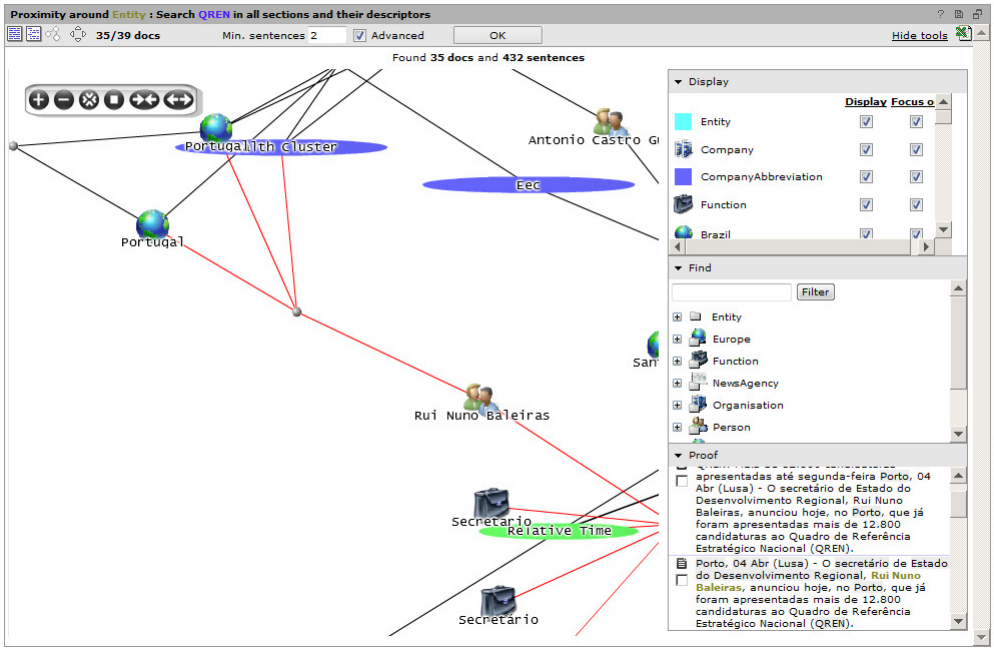

Figura 32 – Detalhe da análise de proximidade

Com base nestes elementos de texto estruturado e anotado, tornou-se possível desenvolver uma série de análises capazes de possibilitar a valorização da informação recolhida, ou seja, conseguiu-se extrair conhecimento a partir da informação textual pública, no presente caso, por intermédio destes processos.

Através destas funcionalidades, o modo como o analista de informação desempenha as suas análises inverte-se. Ao invés de pesquisar textos e de os ler para posteriormente tentar estabelecer as relações entre as entidades, o analista parte de uma grande *"árvore de relações"* que vai "*podando"* de acordo com os propósitos que pretende alcançar. Os testes efectuados pelos utilizadores desta solução confirmam que o processo de análise de conteúdos é significativamente mais rápido, simples e eficaz que o tradicional baseado na pesquisa e leitura dos textos.

A análise preliminar da aplicação destas ferramentas de estruturação de texto sobre os processos de trabalho da Manchete permitiu verificar que há ganhos operacionais e estratégicos. Em termos operacionais, destacou-se a rapidez e uniformização dos critérios de classificação das notícias, o que proporcionou o

aumento do nível do serviço prestado ao cliente. No contexto estratégico, estas funcionalidades têm permitido o desenvolvimento de uma plataforma de conhecimento como base para a criação de novos produtos e serviços, como é o caso do desenvolvimento de *skill cartridges* para sectores de actividade específicos.

De modo a complementar a apresentação dos resultados obtidos, apresenta-se de seguida um exemplo de utilização da aplicação Luxid com *skill cartridge* TM 360 português na óptica da IC.

5.2. Exemplo de utilização na óptica da IC

No presente exemplo pretende-se obter conhecimento com base na informação textual das notícias, minimizando o recurso à sua leitura através da utilização dos recursos de análise que a aplicação Luxid conjuntamente com o *skill cartridge* TM 360 português disponibiliza aos utilizadores.

Para a execução deste exemplo, foram inseridas no sistema 23.795 notícias relacionadas com temas de grande consumo. Dessas, 15.897 referem-se ao ano de 2009 sendo as restantes publicadas até Agosto de 2010.

Apresentam-se de seguida as etapas realizadas.

5.2.1. Definição da pesquisa

Efectuou-se uma pesquisa dos documentos que contenham a palavra “leite”, confinada ao tema “distribuição”. Não foi aplicado nenhum filtro temporal sobre os documentos, não foi seleccionada nenhuma entidade em particular nem se aplicaram quaisquer outros filtros sobre os metadados inerentes aos documentos carregados em sistema.

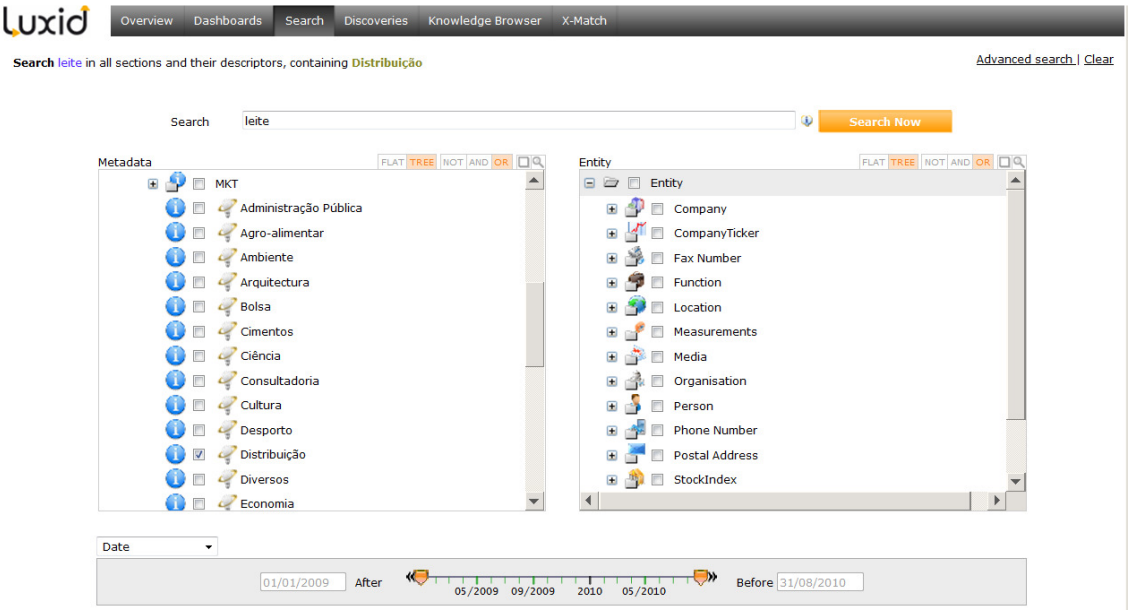

Figura 33 – Definição da pesquisa por *leite* e *distribuição*

5.2.2. Resultado da pesquisa

A pesquisa retornou um conjunto com 172 documentos. As notícias cobrem uma vasta área de publicações, desde revistas culinárias até revistas de saúde, passando pelos jornais de grande distribuição.

Optou-se por trabalhar sobre este conjunto de notícias, não se filtrando os órgãos de comunicação social donde estas provinham.

Figura 34 – Lista dos documentos resultantes da pesquisa

5.2.3. Análise do conjunto de notícias

Procedeu-se de seguida à análise de cruzamento de entidades que permitisse vislumbrar relações entre pessoas e empresas (incluí marcas de produtos).

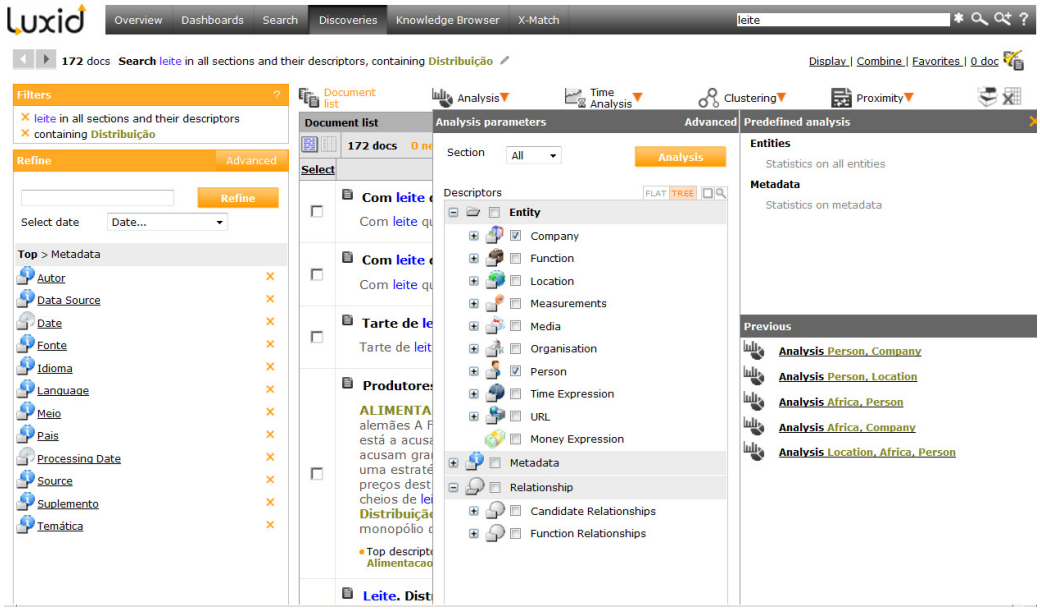

Figura 35 – Selecção de parâmetros para a análise de cruzamento de entidades

5.2.4. Tabela de resultados: Pessoas Vs. Empresas

A tabela de resultados permite vislumbrar uma relação forte entre a empresa Nestlé e Grace Mugabe.

Ao deparar-se com esta relação, um analista de IC poderia ter a curiosidade em responder às seguintes questões:

- Quem é Grace Mugabe?
- Qual a sua relação com a Nestlé?

Na tentativa de dar resposta às questões anteriores, o analista de IC poderia ter a apetência para ler os documentos onde estas duas entidades surgem. Contudo, como a leitura dessas notícias poderá ser longa e improdutiva, levando a eventuais perdas de tempo, optou-se por não seleccionar e ler os documentos que o sistema utilizou para estabelecer esta relação.

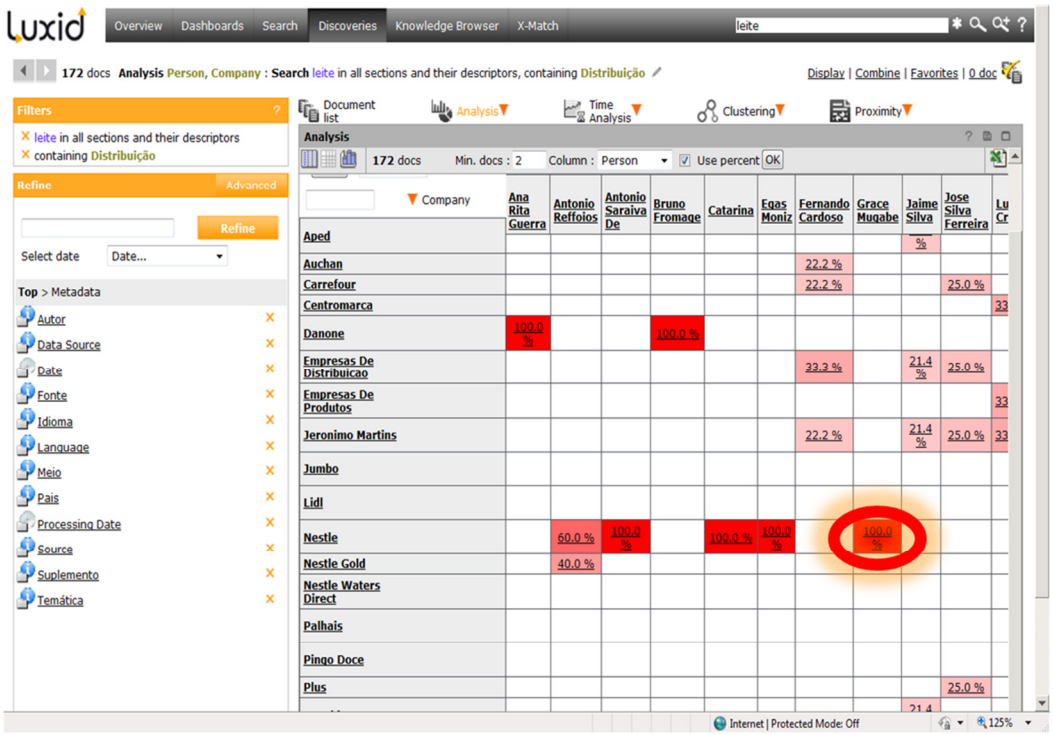

Figura 36 – Tabela de resultados da análise de cruzamento de entidades

Neste contexto, optou-se por aplicar uma análise de *cluster* sobre os documentos, onde estes são agrupados por termos ou expressões que tenham em comum.

5.2.5. Clustering

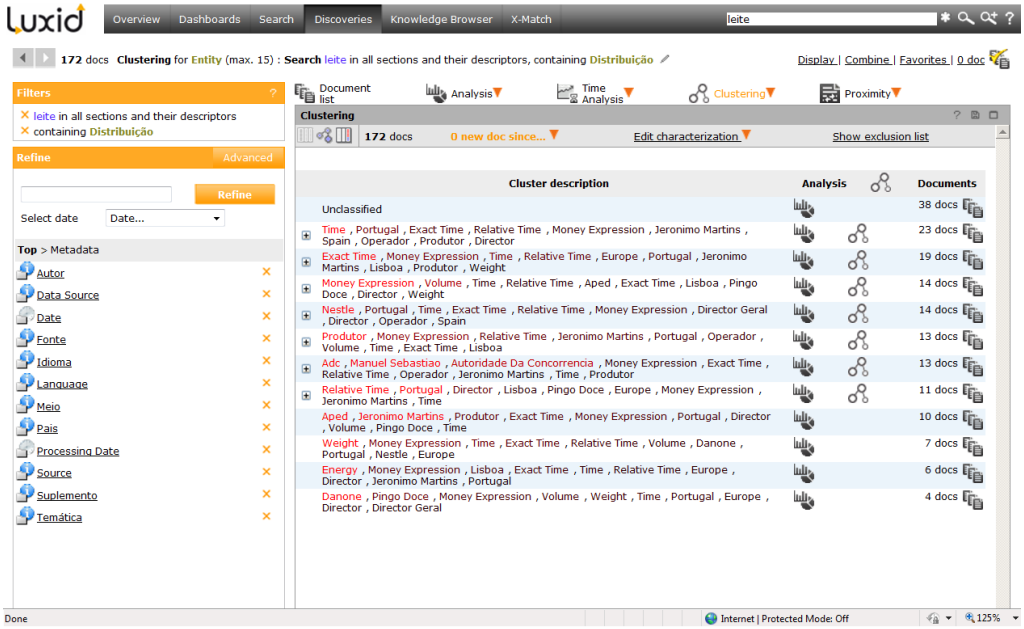

Figura 37 – Lista de *clusters*

Os documentos foram distribuídos em 11 grupos, havendo um 12º grupo que agrega as notícias que não foram possíveis associar a nenhum dos restantes grupos por não partilharem termos ou expressões semelhantes.

O grupo 4 agrega 14 documentos que têm como denominador comum a Nestlé. O peso do termo "Nestlé" detém neste agrupamento situa-se na ordem do 94%.

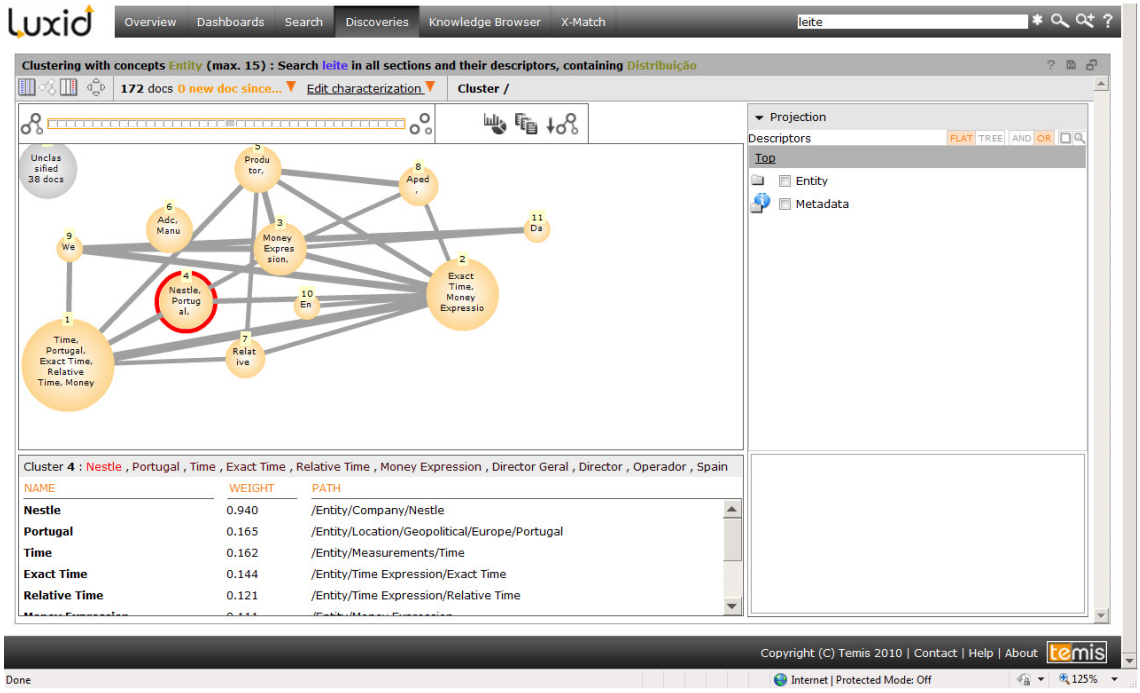

Figura 38 – Representação gráfica dos *clusters*

Nesta fase, podia-se optar por continuar a análise apenas sobre os 14 documentos ou permanecer com o conjunto de 172 notícias que resultaram da pesquisa inicial. Optou-se por continuar com o maior número de documentos e avançar para uma análise gráfica de proximidade entre as entidades.

5.2.6. Análise de proximidade entre entidades

Efectuou-se uma análise de proximidade entre entidades. Uma relação de proximidade ocorre quando duas ou mais entidades surgem muito próximas, como na mesma frase, por exemplo.

A análise de proximidade retornou diagrama de relações que se apresenta na imagem seguinte.

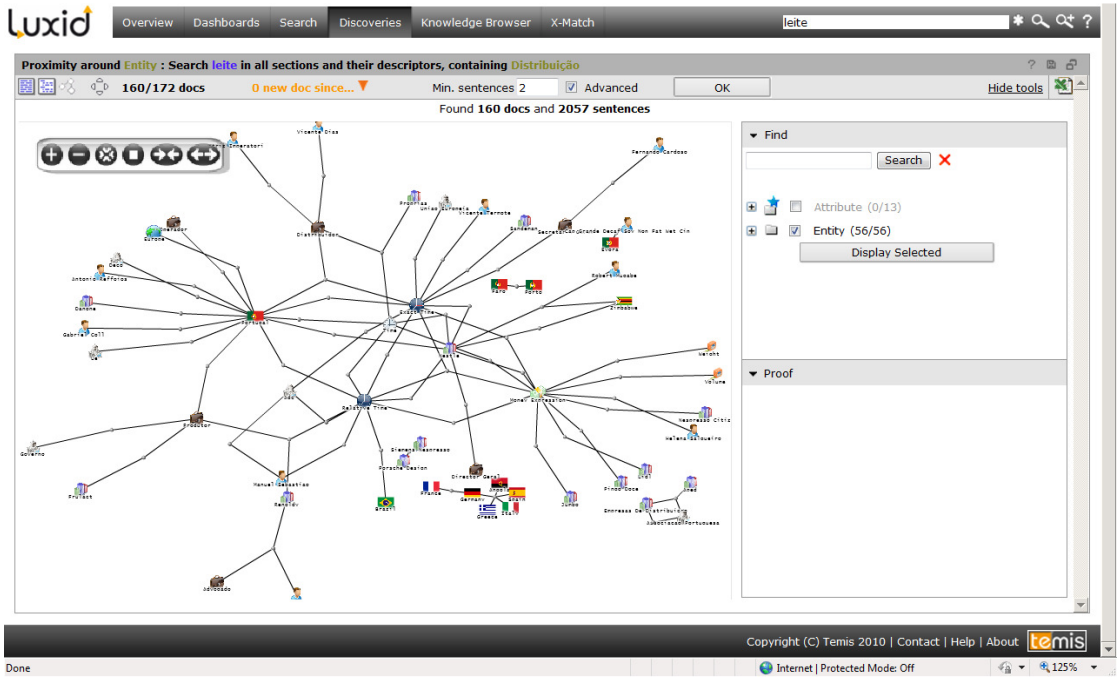

Figura 39 – Panorâmica da análise de proximidade

Seleccionado a entidade que representa a pessoa “Robert Mugabe”, obteve-se o seguinte detalhe.

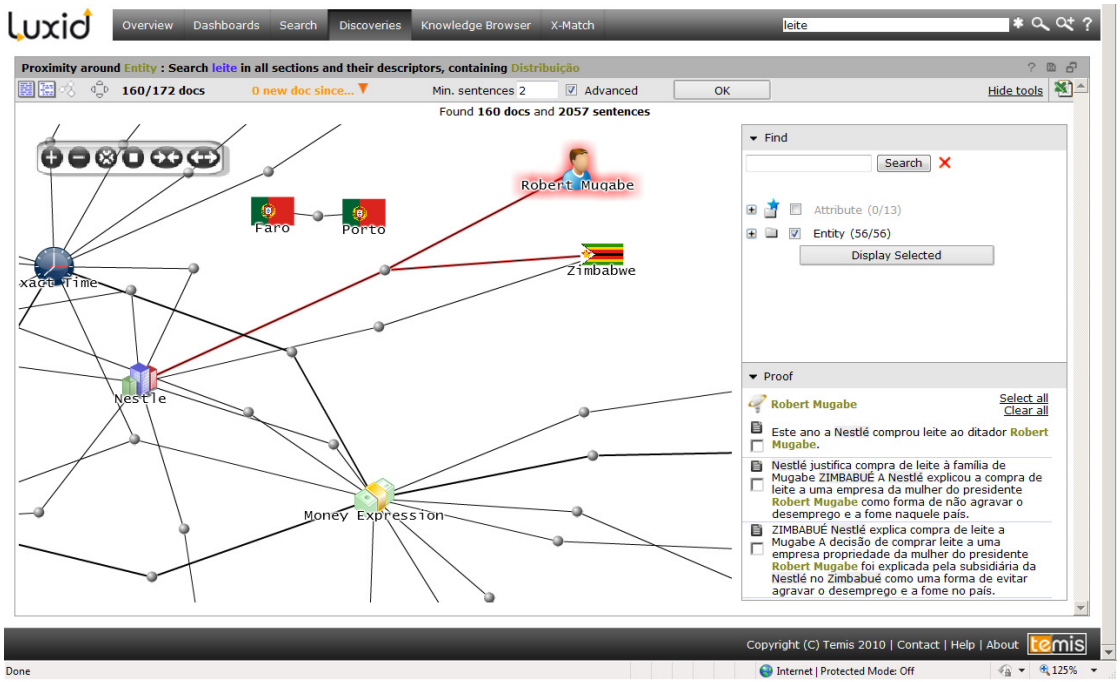

Figura 40 – Detalhe da análise de proximidade sobre Robert Mugabe

A imagem apresenta uma relação de proximidade evidente entre Robert Mugabe e a Nestlé. No canto inferior direito da imagem surge a prova deste relacionamento. Sem ser necessário ler integralmente as notícias, consegue-se obter o seguinte conhecimento sobre os factos que relacionam a Nestlé e Robert Mugabe:

- "Este ano a Nestlé comprou leite ao ditador **Robert Mugabe**."
- "Nestlé justifica compra de leite à família de Mugabe ZIMBABUÉ A Nestlé explicou a compra de leite a uma empresa da mulher do presidente **Robert Mugabe** como forma de não agravar o desemprego e a fome naquele país."
- "ZIMBABUÉ Nestlé explica compra de leite a Mugabe A decisão de comprar leite a uma empresa propriedade da mulher do presidente **Robert Mugabe** foi explicada pela subsidiária da Nestlé no Zimbabué como uma forma de evitar agravar o desemprego e a fome no país."

Nesta fase, já se torna possível vislumbrar de forma nítida qual a relação mantida entre a Nestlé e Grace Mugabe. Apesar de já se saber quem é Grace Mugabe vale a pena tentar descobrir mais conhecimento sobre a relação estabelecida entre a Nestlé e o regime de Mugabe.

Seleccionado o símbolo representado pela bandeira do Zimbabué, são revelados os seguintes factos:

- "- Aqui em Portugal não houve qualquer reacção à questão do **Zimbabué**."
- Os partidários de Mugabe acusam a Nestlé de impor sanções à "primeira família do **Zimbabué**" e alguns deslocaram-se mesmo à empresa para exigir a manutenção do negócio do leite."
- "Nestlé decide parar no **Zimbabué"**
- "POLÉMICA COM GRACE MUGABE Nestlé decide parar no **Zimbabué** A Nestlé suspendeu as suas actividades no **Zimbabué**, depois

de ter recebido ameaças relacionadas com a decisão de deixar de comprar leite à fazenda da primeira-dama do país africano, Grace Mugabe, anunciou o porta-voz da empresa no Quénia, em declarações ao diário oficial zimbabueano The Herald."

- A aquisição de mais de um milhão de litros de leite anuais a Grace Mugabe levanta questões éticas, não só pela forma como os detentores do poder no **Zimbabué** se apoderaram das terras de mais de 4500 agricultores, mas também por Mugabe ser alvo de sanções económicas por parte da União Europeia e dos EUA."
- "Nestlé justifica compra de leite à família de Mugabe **ZIMBABUÉ** A Nestlé explicou a compra de leite a uma empresa da mulher do presidente Robert Mugabe como forma de não agravar o desemprego e a fome naquele país."
- "ZIMBABUÉ Nestlé explica compra de leite a Mugabe A decisão de comprar leite a uma empresa propriedade da mulher do presidente Robert Mugabe foi explicada pela subsidiária da Nestlé no **Zimbabué** como uma forma de evitar agravar o desemprego e a fome no país."

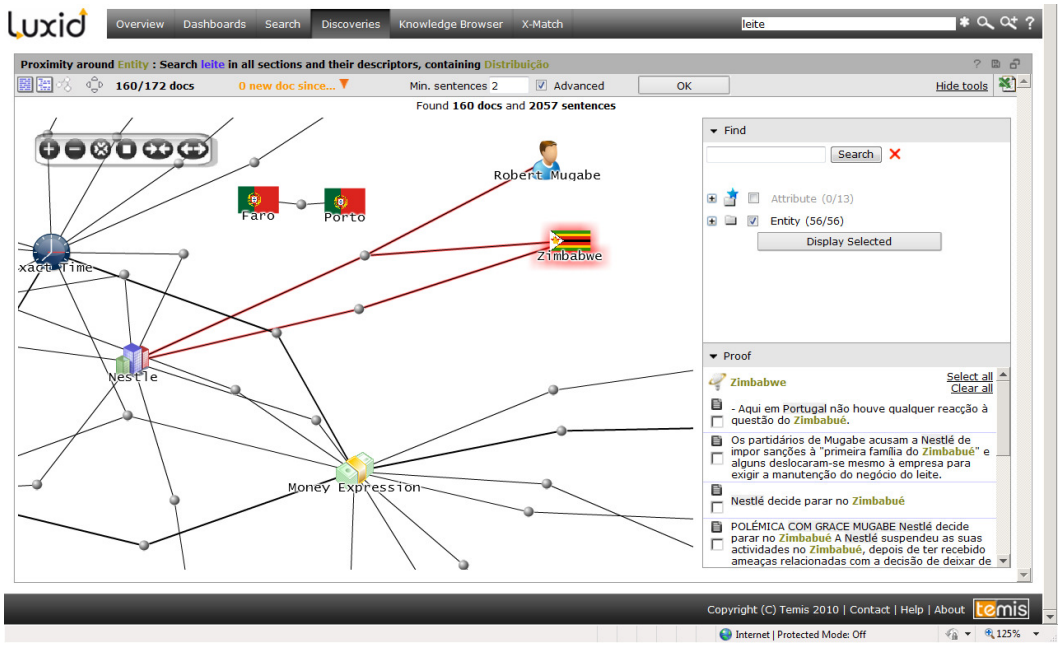

Figura 41 – Detalhe da análise de proximidade sobre Zimbabué

Com os elementos recolhidos até ao momento, não só se conseguiu responder às duas questões colocadas como foi possível saber que a quantidade de leite comercializado rondava o milhão de litros por ano e que a Nestlé deixou de comprar leite à família Mugabe. Soube-se também dos argumentos apresentados pela Nestlé para justificar este negócio, apesar do embargo internacional decretado pela União Europeia e Estados Unidos da América ao Zimbabué, da não reacção portuguesa a este negócio e das ameaças recebidas pela Nestlé.

Nesta fase do estudo, já se torna viável ler alguns dos documentos que sustentam o conhecimento adquirido ao longo da análise, apesar dessa leitura não ser indispensável.

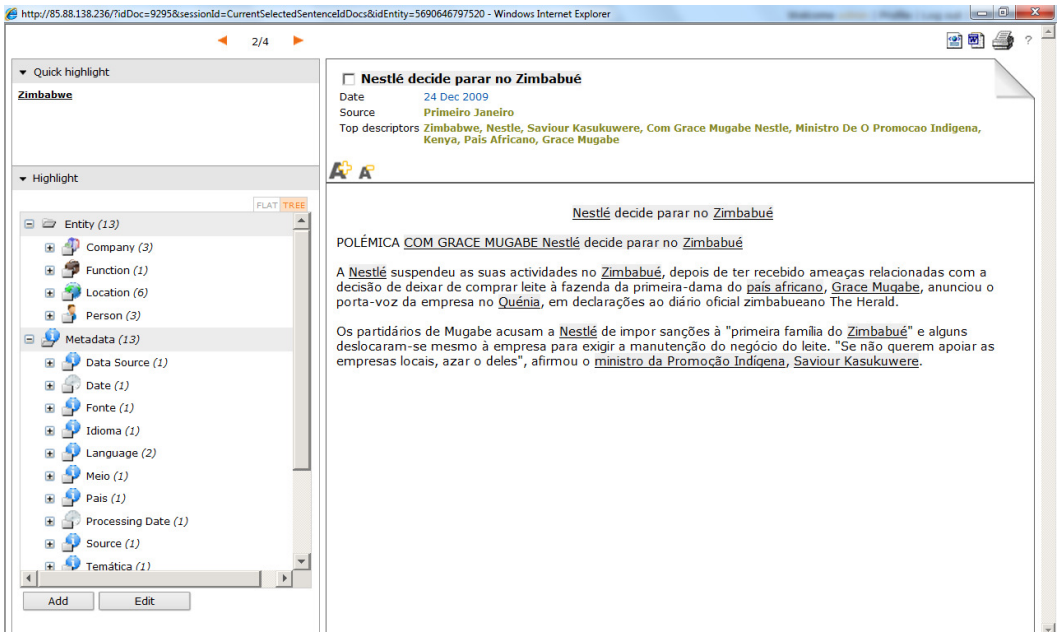

Figura 42 – Notícia do Primeiro de Janeiro relativa ao assunto

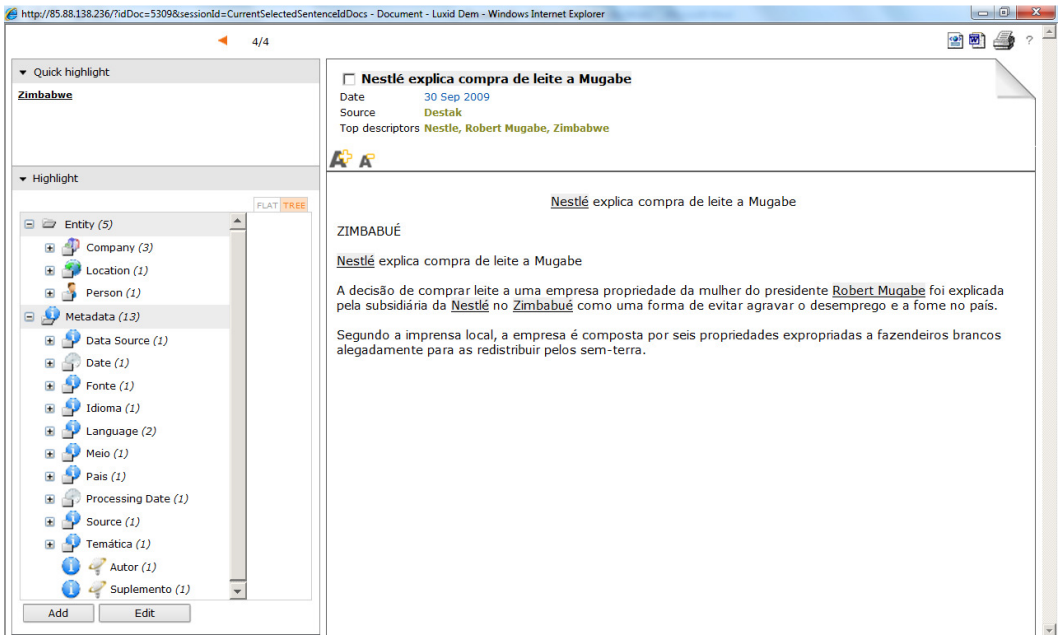

Figura 43 – Notícia do Destak relativa ao assunto

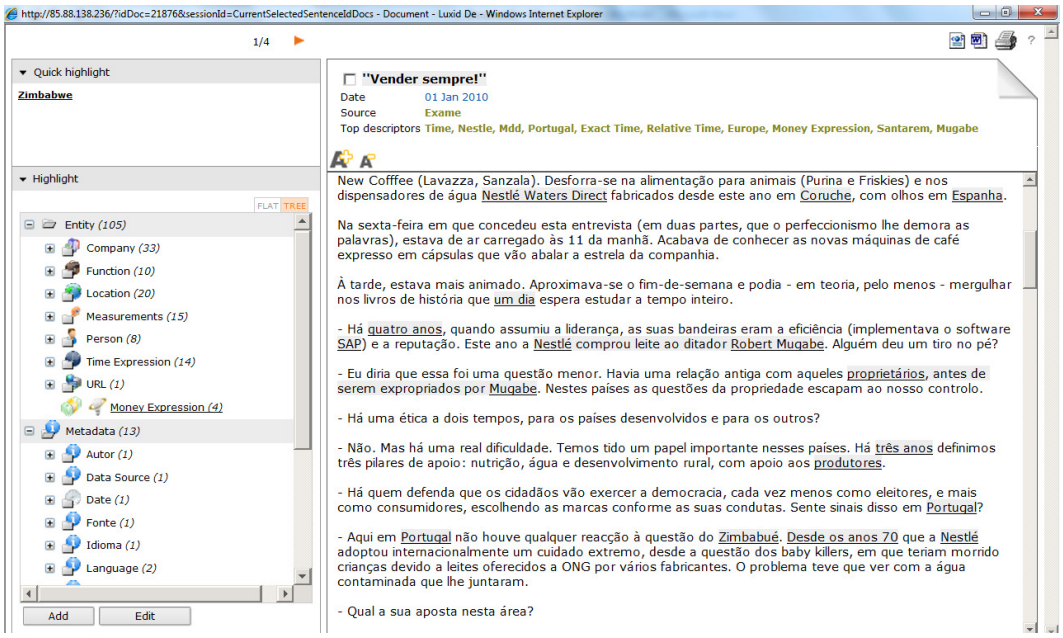

Figura 44 – Entrevista da revista Exame ao director-geral da Nestlé em Portugal

Apesar das questões levantadas ao longo do estudo estarem respondidas, optou-se por realizar uma análise mais profunda sobre o cluster 4 referido anteriormente,

isto é, o conjunto de documentos onde a Nestlé surgia como elemento mais relevante.

5.2.7. Knowledge Browser

Sobre o conjunto de 14 documentos do cluster 4, aplicou-se a análise gráfica *Knowledge Browser*. O *Knowledge Browser* difere da análise de proximidade por estabelecer como critério de relacionamento entre entidades a existência de uma acção (verbo) ao invés de recorrer ao critério da proximidade das palavras. A acção que provoca o relacionamento entre relações é apresentada no diagrama.

Neste caso, foi possível identificar um relacionamento entre entidades que envolvia um valor monetário. Seleccionado esse valor, foi possível verificar que a Nestlé investiu três milhões de euros na sua fábrica de Avança, tendo em conta as seguintes frases:

- "Nestlé Portugal estima duplicar, na fábrica de Avança, a produção de cereais que têm como destino os mercados de exportação, graças ao investimento de três milhões de euros"
- "Nestlé Portugal estima duplicar, na fábrica de Avança, a produção de cereais que têm como destino os mercados de exportação, graças ao investimento de três milhões de euros"

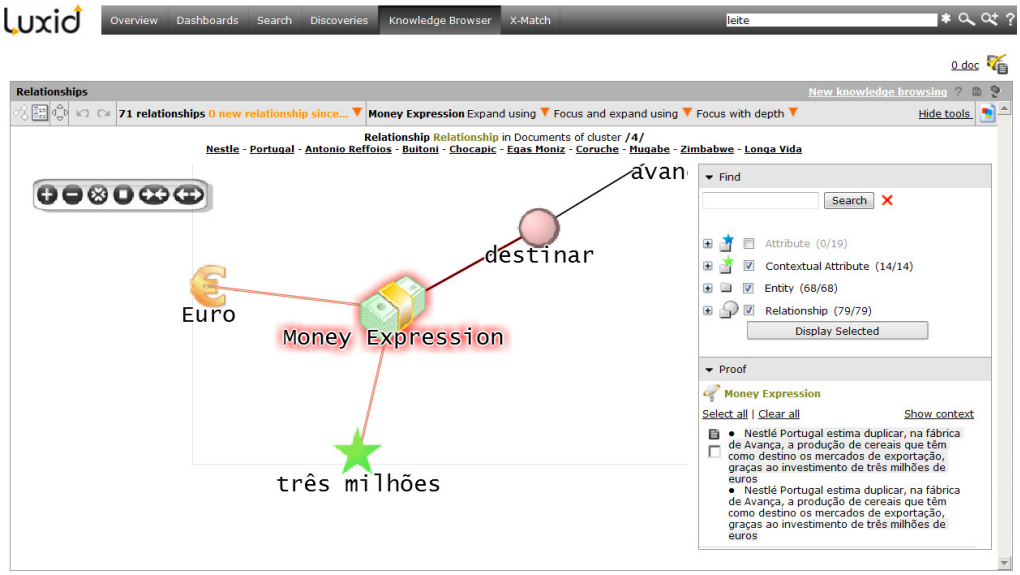

Figura 45 – Detalhe do Knowledge Browser relativo ao montante de investimento

5.2.8. Conclusões do exemplo

Este exemplo mostrou as capacidades analíticas oferecidas pela solução Luxid com a utilização do *skill cartridge* TM 360 português sobre um vasto conjunto de notícias. Num curto período de tempo, um analista de informação ou de inteligência competitiva foi capaz de focar a sua análise nos aspectos essenciais, descartando informação pouco relevante e obter conhecimento valioso para os propósitos da sua análise.

6. Conclusões e desenvolvimentos futuros

Para uma empresa cujo modelo de negócio assenta na gestão e exploração de grandes quantidades de informação textual como é o caso da Manchete, as conclusões resultantes do trabalho descrito perspectivam uma mudança não só na sua forma de trabalhar, mas também no seu *core business*. Até agora, as empresas de gestão de informação como a Manchete têm o seu modelo de negócio na órbita dos serviços de *clipping*, ou seja, focado na recolha de dados e na criação de repositórios onde estes são armazenados e trabalhados na perspectiva de obter alguma informação que gere valor ao ser disseminada pelos seus clientes. Com o desenvolvimento deste *skill cartridge* estão criadas as condições para que a empresa deixe de ter a disponibilização de conteúdos (*content provider*) como principal área de negócio, para se assumir como um especialista na descoberta e gestão do conhecimento. No caso particular da Manchete, é espectável que o *skill cartridge* agora apresentado tenha impacto significativo na empresa, quer na vertente operacional quer na redefinição da sua estratégia de negócio.

Em termos operacionais, a adopção destas soluções analíticas de texto proporcionam uma sistematização e automatização dos processos de *clipping* e uma redefinição do modo como os relatórios de inteligência competitiva são elaborados. Os ganhos no *clipping* resultam da criação de modelos de categorização que permitem organizar as notícias por temas. Para além de uniformizar os critérios de categorização, há que entrar em linha de conta com a maior rapidez no processamento da informação, sendo esta uma questão crítica na medida que o tempo disponível para o tratamento da informação desde que é recebida até ser remetida ao cliente dever ser tão breve quanto possível. Já na perspectiva dos relatórios de inteligência competitiva, os ganhos advêm da estruturação e extracção das entidades e representação gráfica das suas relações. Neste caso, o processo de investigação tornar-se mais simples e rápido, para além de permitir agregar informação de um conjunto de fontes mais vasto e disperso.

Por outro lado, a criação do *skill cartridge* gerou *know-how* numa área de conhecimento ainda raro em Portugal que merece ser capitalizado, o que propicia um alargamento do negócio da empresa e o reposicionamento desta no mercado. Nesta óptica, é natural que a empresa consolide uma área de negócio dedicada à disponibilização deste conhecimento às empresas e organizações que sintam necessidade em analisar informação interna não estruturada.

No sentido de consolidar ainda mais esta estratégia de mudança, os desenvolvimentos futuros deste sistema recairão sobre as restantes duas etapas do plano global apresentado nas secções anteriores, ou seja, na análise de relações e na análise dos sentimentos.

Apesar de se assemelhar à análise de proximidade apresentada anteriormente, a análise de relações através da qual se desenvolve a operacionalização do conhecimento, distingue-se por identificar as acções através das quais essas relações são estabelecidas, não se confinando a verificar se as entidades surgem próximas no texto.

Actualmente, as relações entre entidades são estabelecidas de forma excessivamente lata, em virtude da ocorrência de um verbo (que indicia acção) entre duas entidades, ser condição suficiente para o estabelecer de uma relação. Deste modo, deverão ser criados *skill cartridges* específicos para cada área de interesse, com a capacidade de filtrarem as entidades relacionáveis bem como os motivos que originam o estabelecer das relações. Deste modo, obter-se-á uma ferramenta analítica ainda mais assertiva e focada na descoberta de conhecimento accionável para o sector de actividade em questão.

Um exemplo desta análise é apresentado na figura 46.

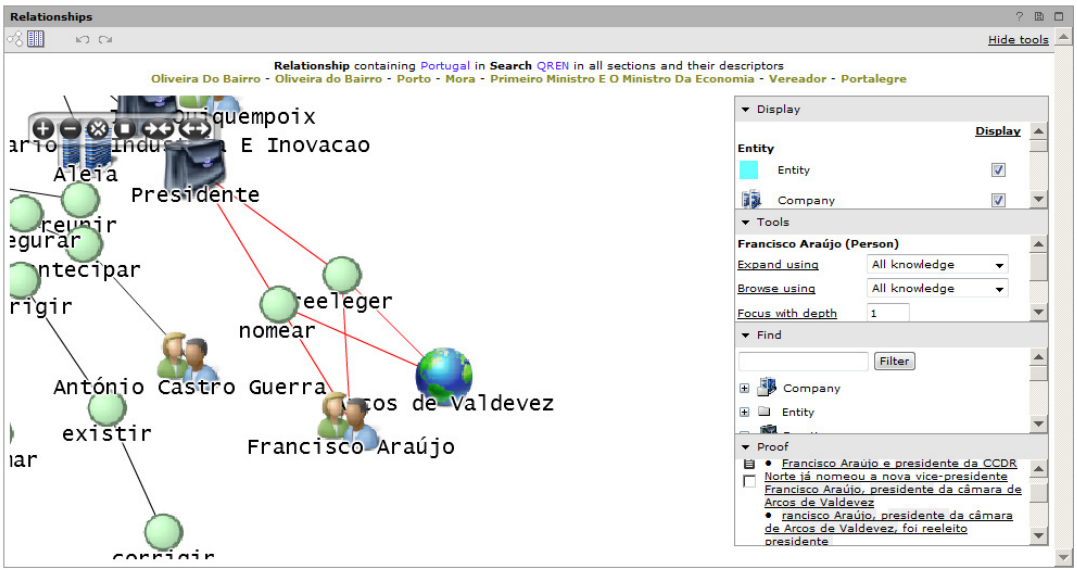

Figura 46 – Exemplo da análise de relações

Por seu turno, a análise de sentimentos incidirá sobre as expressões qualificativas usadas para caracterizar a aceitação ou recusa sobre determinado produto ou serviço, previamente extraído por intermédio dos processos descritos nos capítulos anteriores. A análise de sentimentos permite a segmentação do contexto de análise nas categorias de favorabilidade de sentimento, expectativas e riscos, de acordo com a sua natureza e intensidade:

Para cada um destes tipos, poder-se-á classificar a tipologia da análise de acordo com a seguinte tabela.

Sentimentos Positivos	**Sentimentos Negativos**	**Riscos**	**Expectativas**
Positivo	Negativo	Negócio	Grande Expectativa
Muito Positivo	Muito Negativo	Legais	Pequena Expectativa

Tabela 5 – Tipologia da Análise de Sentimentos

Exemplos típicos da utilidade da análise de sentimentos são a verificação do impacto de campanhas de Marketing, recolha de feedback dos clientes relativamente a produtos e serviços, vislumbrar de novas tendências de mercado e identificação de riscos para o negócio. A figura 47 disponibiliza algumas das aplicações da análise de sentimentos por tipologia.

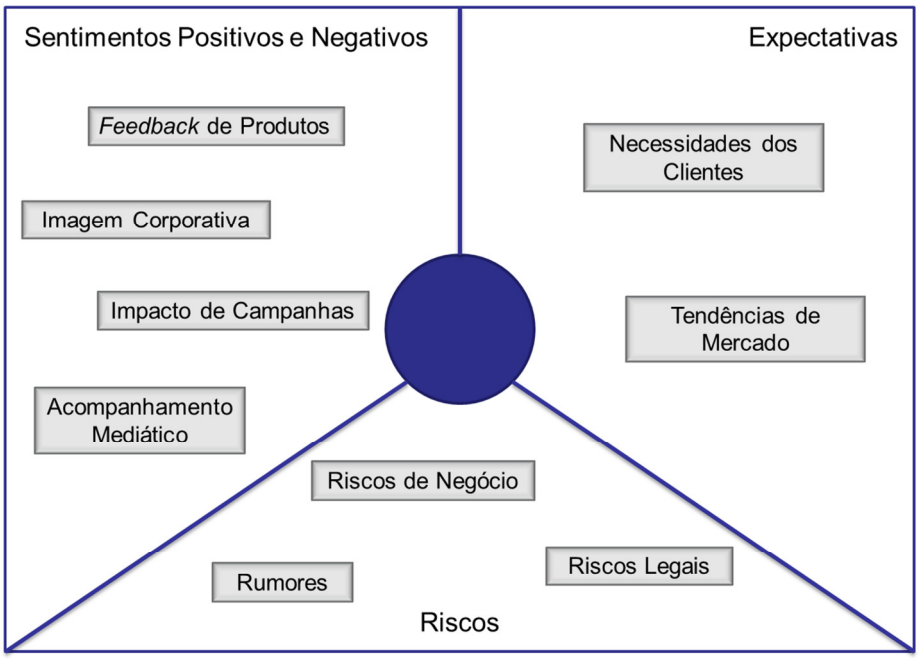

Figura 47 – Exemplos de aplicação da análise de sentimentos por tipologia

A figura 48 apresenta um exemplo da utilização da análise de sentimentos no contexto de captar os principais aspectos positivos e negativos relativos às funcionalidades do iPhone.

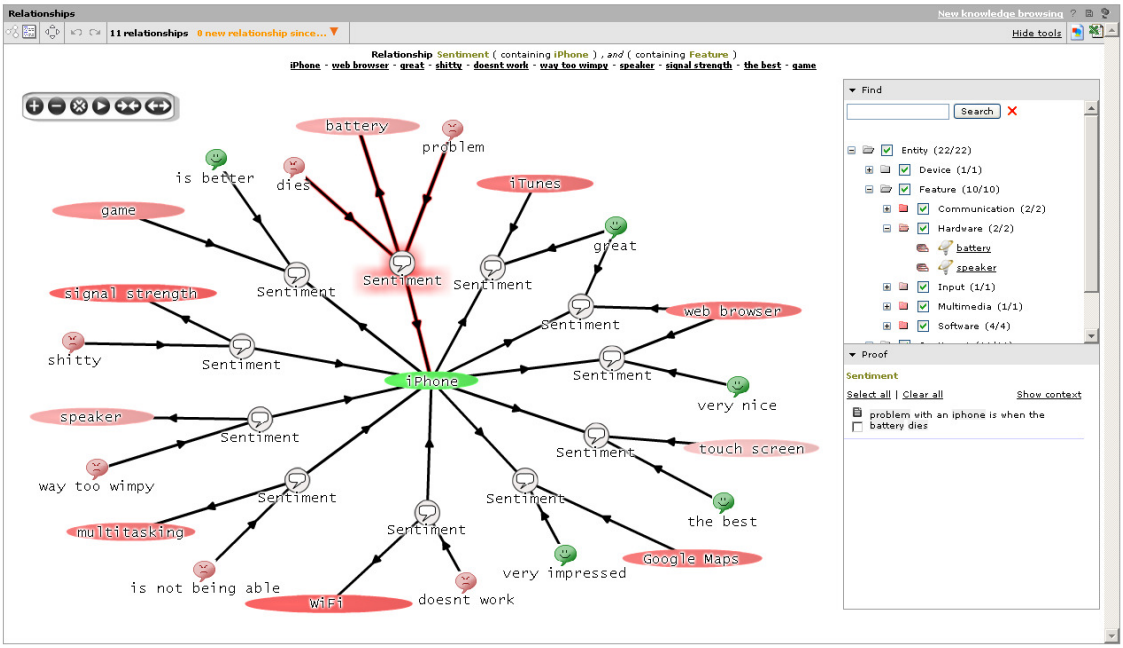

Figura 48 – Representação dos sentimentos associados a um produto

Uma referência final sobre o âmbito dos temas abordados nesta dissertação. Apesar do projecto se focar no desenvolvimento e aplicabilidade de técnicas de text mining a uma empresa de gestão de informação, a sua utilização não se confina a este meio. Este trabalho evidencia a importância dos processos de text mining num ambiente económico mais vasto e onde a informação flui livremente.

Deseja-se que este vasto e complexo plano não ensine os computadores a ler mas que os ajude a interpretar textos escritos na língua portuguesa e permita obter valor dos seus conteúdos. Ter-se-á então contribuído de forma decisiva para sistematizar o processo de converter dados dispersos em informação e desta extrair valor donde resultará o conhecimento.

Referências Bibliográficas

About SCIP. (n.d.). Retrieved 13-07-2011, from http://www.scip.org/content.cfm?itemnumber=2214&navItemNumber=492

ACI. (n.d.). Retrieved 13-07-2011, from http://www.academyci.com

AMEC. (n.d.). Retrieved 13-07-2011, from http://www.amecorg.com/amec/index.asp

Ansoff, H. I. (1980). Strategic issue management. *Strategic Management Journal*, 1(2), 131-148. doi: 10.1002/smj.4250010204

Bellinger, G., Castro, D., & Mills, A. (2004). Data, Information, Knowledge, and Wisdom Retrieved 13-07-2011, from http://www.systems-thinking.org/dikw/dikw.htm

Blumberg, R., & Atre, S. (2003). The Problem with Unstructured Data. Retrieved 13-07-2011, from http://soquelgroup.com/Articles/dmreview_0203_problem.pdf

CAPSI. (n.d.). Retrieved 13-07-2011, from http://www.estv.ipv.pt/dep/di/capsi2009

Carlisle, J. P. (2007). A Look into the Relationship between Knowledge Management and the Knowledge Hierarchies. *System Sciences*, 2007. HICSS 2007. 40th Annual Hawaii International Conference on, 183a-183a.

Coimbra, J. (2009, 19/11/2009). Enterprise Content Management Market and Trends. Paper presented at the *Como Garantir a Conformidade e Acelerar o "Time to Market" num Contexto Exigente?*, Centro Cultural de Belém, Lisboa.

Conferência Informação Estratégica e Inovação. (n.d.). Retrieved 13-07-2011, from http://www.mynetpress.com/conferenciafuturo

Cornish, S. L. (1997). Product Innovation and the Spatial Dynamics of Market Intelligence: Does Proximity to Markets Matter? *Economic Geography*, 73(2), 143-165.

Courtney, J. F. (2001). Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for DSS. *Decision Support Systems*, 31(1), 17-38.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707-1720. doi: DOI 10.1016/j.eswa.2007.01.035

Domenig, R., & Dittrich, K. R. (1999). An overview and classification of mediated query systems. *SIGMOD Rec.*, 28(3), 63-72. doi: http://doi.acm.org/10.1145/333607.333615

European and Arabic Linguistic Suite. (n.d.). Retrieved 13-07-2011, from http://www.teragram.com/oem/euro_lang.htm

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview (pp. 18): *American Association for Artificial Intelligence*.

Ferrari, M. (2005). ROI in text mining projects. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

FIBEP. (n.d.). Retrieved 13-07-2011, from http://www.fibep.info/fibep/en

Fleisher, C. S. (2008). Using open source data in developing competitive and marketing intelligence. [Article]. *European Journal of Marketing*, 42(7-8), 852-866. doi: 10.1108/03090560810877196

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1), 70-105. doi: 10.1108/EUM0000000007162

Gantz, J., & Reinsel, D. (2010). The Digital Universe Decade – Are You Ready? Retrieved 13-07-2011, from http://idcdocserv.com/925

Gao, L., Chang, E., & Han, S. (2005). Powerful Tool to Expand Business Intelligence: Text Mining. Paper presented at the *PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY*, VOL 8 8: 110-115 2005.

Hearst, M. A. (1999). Untangling text data mining. 3-10.

Hey, J. (2004). The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link: *Intergovernmental Oceanographic Commission (UNESCO)*.

Hobbs, J. R. (1993). The generic information extraction system. Paper presented at the Proceedings of the *5th conference on Message understanding*, Baltimore, Maryland.

Kosovac, B., Froese, T. M., & Vanier, D. J. (2000). Integrating Heterogeneous Data Representations in Model-Based AEC/FM Systems. Paper presented at the *Construction Information Technology 2000, Proceedings of CIT 2000* – The CIB-W78, IABSE Vol. 2, pp 556-567.

Kroeze, J. H., Matthee, M. C., & Bothma, T. J. D. (2003). Differentiating data- and text-mining terminology. Paper presented at the *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*.

Li, G., Ooi, B. C., Feng, J., Wang, J., & Zhou, L. (2008). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. Paper presented at the Proceedings of the *2008 ACM SIGMOD international conference on Management of data*, Vancouver, Canada.

Luken, M., & Baisch, F. (1998). Decision support based on weak signals-overcoming the implementation gap of strategic early warning systems. *Systems, Man, and Cybernetics*, 1998. 1998 IEEE International Conference on, 3, 2484-2488 vol.2483.

Luxid 5.0 User Guide. (n.d.). (2008). France: Temis.

Maguitman, A. G., Menczer, F., Erdinc, F., Roinestad, H., & Vespignani, A. (2006). Algorithmic computation and approximation of semantic similarity. *World Wide Web-Internet and Web Information Systems*, 9(4), 431-456. doi: DOI 10.1007/s11280-006-8562-2

Mannila, H. (2000). Theoretical frameworks for data mining. *SIGKDD Explor. Newsl.*, 1(2), 30-32. doi: http://doi.acm.org/10.1145/846183.846191

Mastering New Challenges in Text Analytics. (n.d.). (2008) Retrieved 13-07-2011, from http://www.spss.com/media/collateral/MCTWP-0408lr.pdf

McKellar, H. (2009). KMWorld 100 Companies That Matter in Knowledge Management Retrieved 25-08-2010, from http://www.kmworld.com/Articles/Editorial/Feature/KMWorld-100-Companies-That-Matter-in-Knowledge-Management-52787.aspx

McKnight, W. (2005). Text Data Mining in Business Intelligence. Information Management Magazine Retrieved 13-07-2011, from http://www.information-management.com/issues/20050101/1016487-1.html

Milic-Frayling, N. (2005). Text processing and information retrieval. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems journal*, 40(4), 967-984.

Nurnberger, A., Seising, R., & Wenzel, C. (2009). On the fuzzy interrelationships of data, information, knowledge and wisdom. *Fuzzy Information Processing Society, 2009. NAFIPS 2009*. Annual Meeting of the North American, 1-6.

Pazienza, M. T. (2005). Information extracting and... surroundings. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

Peters, G. (2005). Media industry: How to improve documentalists efficiency. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

Plejic, B., Vujnovic, B., & Penco, R. (2008, 21-22 Dec. 2008). Transforming unstructured data from scattered sources into knowledge. Paper presented at the *Knowledge Acquisition and Modeling Workshop*, 2008. KAM Workshop 2008. IEEE International Symposium on.

Prado, H. A. d., & Ferneda, E. (Eds.). (2007). *Emerging Technologies of Text Mining: Techniques and Appllications*. Hersey, New York: Information Science Reference.

Rebelo, F. (2009, 16-02-2009). [Media, Reputação e Inteligência].

Schanz, K.-U. (2006). Reputation and Reputational Risk Management. Geneva Papers on *Risk & Insurance - Issues & Practice*, 31(3), 377-381.

SCIP. (n.d.). Retrieved 13-07-2011, from http://www.scip.org

SIIA. (n.d.). Retrieved 13-07-2011, from http://www.siia.net/codies/2010/winners.asp#content

Skill Cartridge Author's Guide. (n.d.). (2010). France: Temis.

Smullen, C. W., Tarapore, S. R., & Gurumurthi, S. (2007). A Benchmark Suite for Unstructured Data Processing. Paper presented at the *The 4th International Workshop on Storage Network Architecture and Parallel I/Os (SNAPI'07)*

in conjunction with the *24th IEEE Conference on Mass Storage Systems and Technologies (MSST)*, San Diego, California, USA. http://www.cs.virginia.edu/~gurumurthi/papers/snapi07.pdf

Stenmark, D. (2002). Information vs. knowledge: the role of intranets in knowledge management. Paper presented at the *System Sciences*, 2002. HICSS. Proceedings of the *35th Annual Hawaii International Conference* on.

Sullivan, D. (2001). Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales. New York, NY, USA: John Wiley & Sons, Inc.

Sullivan, D. (2005). Application integration in applied text mining. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

TM360 Skill Cartridge User Guide. (n.d.). (2007). France: Temis.

Tuomi, I. (1999). Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. Paper presented at the *System Sciences*, 1999. HICSS-32. Proceedings of the *32nd Annual Hawaii International Conference* on.

Uys, J. W., du Preez, N. D., & Uys, E. W. (2008). Leveraging unstructured information using topic modelling. *Management of Engineering & Technology*, 2008. PICMET 2008. Portland International Conference on, 955-961.

Webster, M. (2009). Worldwide Content Management Software and 2008 Vendor Shares Retrieved 09-12-2010, from http://www.idc.com/getdoc.jsp?containerId=219816

White, C. (2005). Consolidating, Accessing and Analyzing Unstructured Data Retrieved 13-07-2011, from http://www.b-eye-network.com/view/2098

XeLDA White Paper. (n.d.). Meylan: Temis.

Zanasi, A. (2005a). Open sources automatic analysis for corporate and governmental intelligence. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

Zanasi, A. (Ed.). (2005b). *Text Mining and its Applications to Intelligence CRM and Knowledge Management*. Southampton: WIT Press.

Anexo I - Portuguese Part-of-Speech Tagset

Tag	Description	Example
+ADJ	invariant adjective	[duas saias] cor-de-rosa
+ADJPL	plural adjective	[cidadaos] portugueses
+ADJSG	singular adjective	[continente] europeu
+ADV	Adverb	directamente
+ADVCOMP	comparison adverb "mais" and "menos"	[um pais] mais [livre]
+AUXBE	finite 'be' ("ser" or "estar")	é, são, estão
+AUXBEINF	infinitive 'be'	ser, estar
+AUXBEINFPRON	infinitive 'be' with clitic	sê-lo
+AUXBEPRON	finite 'be' with clitic	é-lhe
+AUXHAV	finite 'have'	tem, haverá
+AUXHAVINF	infinitive 'have' ("ter", "haver")	ter, haver
+AUXHAVINFPRON	infinitive 'have' with clitic	ter-se
+AUXHAVPRON	finite 'have' with clitic	tinham-se
+CM	comma	,
+CONJ	(coordinating) conjunction	[por fax] ou [correio]
+CONJCOMP	comparison conjunction "do que"	[mais] do que [uma vez]
+CONJSUB	subordination conjunction	para que, se, que
+DEMPL	plural demonstrative	estas
+DEMSG	singular demonstrative	aquele

Tag	Description	Example
+DETINT	interogative or exclamative "que"	[demostra a] que [ponto]
+DETINTPL	plural interogative determiner	quantas [vezes]
+DETINTSG	singular interogative determiner	qual [reacao]
+DETPL	plural definite article	os [maiores aplausos]
+DETRELPL	plural relative determiner	..., cujas [prestações]
+DETRELSG	singular relative determiner	..., cuja [veia poética]
+DETSG	singular definite article	o [serviço]
+DIG	digit	123
+GER	Gerundive	examinando
+GERPRON	gerundive with clitic	deixando-a
+INF	verb infinitive	reunir, conservar
+INFPRON	infinitive with clitic	datar-se
+INTERJ	Interjection	oh, ai, claro
+ITEM	list item marker	A. [Introdução]
+LETTER	isolated character	[da selecção] A
+NEG	Negation	não, nunca
+NOUN	invariant common noun	caos
+NPL	plural common noun	servicos
+NPROP	proper name	PS, Lisboa
+NSG	singular common noun	(Maguitman, Menczer, Erdinc, Roinestad, & Vespignani) rede

+POSSPL	plural possessive	seus [investigadores]
+POSSSG	singular possessive	sua [sobrinha]
Tag	**Description**	**Example**
+PREP	preposition	para, de, com
+PREPADV	preposition + adverb	[venho] daqui
+PREPDEMPL	preposition + plural demonstrative	desses [recursos]
+PREPDEMSG	preposition + singular demonstrative	nesta [placa]
+PREPDETPL	preposition + plural determiner	dos [Grandes Bancos]
+PREPDETSG	preposition + singular determiner	na [construcao]
+PREPPRON	preposition + pronoun	[atras] dela
+PREPQUANTPL	preposition + plural quantifier	nuns [terrenos]
+PREPQUANTSG	preposition + singular quantifier	numa [nuvem]
+PREPREL	preposition + invariant rel. pronoun	[nesta praia] aonde
+PREPRELPL	preposition + plural relative pronoun	[alunos] aos quais
+PREPRELSG	preposition + singular rel. pronoun	[área] através do qual
+PRON	invariant pronoun	se
+PRONPL	plural pronoun	as, eles, os
+PRONSG	singular pronoun	a, ele, ninguém
+PRONREL	invariant relative pronoun	[um ortopedista] que
+PRONRELPL	plural relative pronoun	[as instalações] as quais
+PRONRELSG	singular relative pronoun	[o ensaio] o qual
+PUNCT	other punctuation	: ( ) ;

+QUANTPL	plural quantifier	quinze, alguns, tantos
+QUANTSG	singular quantifier	um, algum, qualquer
Tag	**Description**	**Example**
+SENT	sentence final punctuation	. ! ?
+SYM	symbols	@ %
+VERBF	finite verb form	corresponde
+VERBFPRON	finite verb form with clitic	deu-lhe
+VPP	past participle (also adjectival use)	penetrado, referida