IT Service Management supported by Semantics Intelligence

A Semantic Intelligence Model Applied to an ITIL Implementation

2015

Sérgio Paulo Marques Velho

MESTRADO EM GESTÃO DE SISTEMAS DE INFORMAÇÃO

IT Service Management supported by Semantics Intelligence

Modelo de Inteligência Semântica para uma Implementação ITIL

2015

Sérgio Paulo Marques Velho

Dissertação Orientada pelo Prof. Doutor Mário Macedo

MESTRADO EM GESTÃO DE SISTEMAS DE INFORMAÇÃO

Agradecimentos

O presente estudo resulta não só da vontade pessoal de alcançar mais um grau de estudos académicos, mas sobretudo pela paixão e interesse pelas tecnologias de informação. O estudo foi tomado não como um trabalho, com a conotação a um esforço indesejado, mas como uma oportunidade de evoluir, aplicando o aprendido, num grau de exigência inalcançável pela simples intenção.

Neste estudo fui contínua e prontamente assistido pelo meu orientador de dissertação, Prof. Doutor Mário Macedo, que me encaminhou nos passos a percorrer e meu deu a força necessária para continuar nos momentos de desânimo, a quem eu deixo o meu mais sincero agradecimento.

Não esquecendo que todos somos pessoas, e que é imprescindível termos um pilar de apoio nos momentos mais difíceis, e de partilha da felicidade nos melhores momentos, não posso deixar o meu muito especial agradecimento à minha esposa e filhos que muitas vezes serviram de guia nesta função que cumpri com satisfação e, sinceramente gozo pessoal, garantindo as condições necessárias e imprescindíveis para que tal fosse possível.

Agradeço também ao corpo docente da Universidade Atlântica que tive a oportunidade de conhecer e com os quais tive o privilégio de apreender novos conhecimentos, seguramente úteis na criação deste documento.

Na impossibilidade de referir todas as pessoas que me apoiaram, deixo um Bem Haja a todos!

Resumo

A crescente utilização de recursos suportados por tecnologias de informação e comunicação, aliado à explosão diária de novos conteúdos e informação atribui responsabilidades acrescidas às estruturas de suporte a estas tecnologias. Desta forma, e com a necessidade de acompanhar esta crescente utilização dos recursos, urge garantir um suporte técnico e preventivo adequado à fiabilidade, segurança e eficácia na utilização dos equipamentos e serviços.

É conhecido também o aumento exponencial de dispositivos que se ligam nos dias de hoje às redes informáticas corporativas, e que permitem aceder à informação de novas e variadas formas, o que obriga a um acrescido esforço na manutenção dos requisitos de segurança e mobilidade imprescindíveis à sua utilização.

Paralelamente assiste-se diariamente à necessidade de redução de recursos, quer tecnológicos, quer humanos para promover o suporte destas realidades, pela via financeira ou de contenção de mão-de-obra.

Com base neste paradigma, torna-se evidente a necessidade de repensar a estratégia das TI e otimizar os seus processos de gestão chave. Enquadrando esta problemática à luz da prospeção de informação, identifica-se uma oportunidade de melhoria, fomentando um processo de sistematização e automatização da gestão da informação organizacional promovendo assim mecanismos autónomos para, desta forma, otimizar os processos de suporte, assentando os pressupostos organizativos numa base bem conhecida como a ITIL.

Palavras-chave: inteligência semântica, *text mining*, ITIL, *service desk*, otimização, categorização, pedido, incidente.

Abstract

The increasing use of features supported by information and communication technology combined with the daily outburst of new content and information assigns additional responsibilities to support these technologies infrastructures. Instigated by the need to monitor this growing use of resources, TI is vital to ensure a technical and preventive support to permit a reliable, secure and efficient use of equipment and services.

TI is also a known the nowadays there is an exponential increase of devices that connect to corporate networks, needing to access information in new and different ways, which requires a greater effort in maintaining mobility and security requirements essential to their use.

At the same time we daily witness the need to reduce resources, whether technological or human to guarantee the support to these realities, the financial means or hand labor contention.

Based on this paradigm, TI becomes evident the need to rethink the TI strategy and optimize TI key management processes. Framing this issue supported in information prospecting science, allows to identify an opportunity for improvement, encouraging a process of systematization and automation of organizational information management thus promoting autonomous mechanisms to optimize the support processes, laying the organizational assumptions on ITIL bases.

Keywords: semantic intelligence, text mining, ITIL, service desk, optimization, categorization, request, incident

Índice

1 INTRODUÇÃO 1
1.1 Contexto / Problema 1
1.2 Questão de investigação 2
1.3 Objetivos da investigação 2
1.4 Metodologia de estudo e pesquisa 3
1.5 Organização da dissertação 7
2 REVISÃO DE LITERATURA 9
2.1 ITIL (*Information Technology Infrastructure Library*) 9
2.1.1 Histórico 10
2.1.2 Conceitos e definições 10
2.1.3 Volumes do ITIL 15
2.1.4 Modelos de maturidade 23
2.2 Gestão da Mudança 29
2.2.1 Gestão de Mudança na ITIL 30
2.3 Análise de Conteúdos Semânticos 31
2.3.1 Gestão do conhecimento 34
2.3.2 Processo de descoberta do Conhecimento 35
2.3.3 Recuperação de informação (*IR – Information Retrieval*) 40
2.3.4 Linguística Computacional 42
2.3.5 Preparação de Corpus Textuais 44
2.3.6 Métodos de mineração de texto 52
2.3.7 Pós-processamento do texto 60
3 METODOLOGIA DE INVESTIGAÇÃO 62
3.1 Metodologias de Investigação 62
3.2 Estratégias de Investigação 63
3.2.1 Estudo de Caso 63
3.2.2 Grounded Theory 64
3.2.3 Action Research 64
3.2.4 Secondary Data 65
3.3 Métodos de Investigação 66
3.3.1 Métodos Qualitativos e Quantitativos 66
3.3.2 Justificação da Abordagem Metodológica de Estudo de Caso 66

4 DESENVOLVIMENTO DO MODELO DE INTELIGÊNCIA SEMÂNTICA..68
4.1 Identificação dos Perfis de Serviços de TI Relevantes....68
4.1.1 Impacto da Aplicação da ITIL nas Organizações....71
4.1.2 Análise das Respostas aos Questionários....72
4.2 Pressupostos do Modelo a Conceber....79
4.3 Processos de Inteligência Semântica Aplicáveis....82
4.3.1 *Inputs, Outputs* e Processos do Modelo....83
4.3.2 Metodologias de Inteligência Semântica Aplicáveis....86
5 ESTUDO DE CASO....93
5.1 Caraterização da Organização Objeto de Estudo de Caso....93
5.1.1 Análise das Tecnologias de Informação Existentes no Município....93
5.1.2 Estatísticas de *Service Desk*....93
5.2 Aplicação do Modelo....94
5.2.1 Recolha e Preparação dos Dados....95
5.2.2 Treino do Modelo....96
5.2.3 Teste do Modelo....99
5.3 Resultados Obtidos....101
6 CONCLUSÕES E TRABALHOS FUTUROS....103
6.1 Conclusões....103
6.2 Restrições ao Estudo....104
6.3 Melhorias Futuras....105
7 ANEXOS....115

Índice de Ilustrações

Ilustração 1 - Metodologia Action Research 5
Ilustração 2- Funcionamento de um processo básico segundo o ITIL v3 12
Ilustração 3 - Ciclo de Deming (PDCA) 14
Ilustração 4 - Livros do ITIL: Service Strategy 17
Ilustração 5 - Livros do ITIL: Service Design 19
Ilustração 6 - Livros do ITIL: Service Transition 20
Ilustração 7 - Livros do ITIL: Service Operation 22
Ilustração 8 - Livros do ITIL: Continual Service Improvement 23
Ilustração 9 - Métodos de investigação 67
Ilustração 10 - Fluxo de um pedido ou incidente 80
Ilustração 11 - Diagrama conceptual do modelo 82
Ilustração 12 - Arquitetura Técnica de Funcionamento do Modelo 84
Ilustração 13 - Fases de Implementação do Modelo 85
Ilustração 14 - Fase de pré-processamento de texto no RapidMiner 89
Ilustração 15 - Exemplo do treino do modelo com k-NN (RapidMiner) 91
Ilustração 16 - Processo de treino do modelo (RapidMiner) 97
Ilustração 17 – Treino do modelo (Performance Vector k-NN - Rapid Miner) 98
Ilustração 18 – Treino do modelo (Performance Vector SVM - Rapid Miner) 99
Ilustração 19 – Treino do modelo (Performance Vector Naive-Bayes - RapidMiner) ... 99
Ilustração 20 – Processo de teste (RapidMiner) 100
Ilustração 21 – Teste do modelo (Performance Vector - RapidMiner) 101

Índice de Gráficos

Gráfico 1 - Distribuição de incidentes por principais categorias 70
Gráfico 2 - Respostas ao questionário (questão 1) 73
Gráfico 3 - Respostas ao questionário (questão 2) 74
Gráfico 4 - Respostas ao questionário (questão 3) 75
Gráfico 5 - Respostas ao questionário (questão 4) 79

Índice de Tabelas

Tabela 1 - Características principais do método estudo de caso 4
Tabela 2 - Níveis PMF e descrição 26
Tabela 3- Níveis de análise por Etapas ou por processos 27
Tabela 4 - Vantagens e desvantagens da gestão de mudança na ITIL 31
Tabela 5 - Técnicas de mineração e considerações chave 33
Tabela 6 - Número de pedidos de preenchimento vs número de respostas ao questionário 72
Tabela 7 - Número de questionários respondidos por tipo de organização 73
Tabela 8 – Respostas ao questionário - frequências absolutas (questão 3) 76
Tabela 9 - Respostas ao questionário - frequências relativas (questão 3) 77
Tabela 10 - Respostas ao questionário - Ordenação das categorias acima da média (questão 3) 78
Tabela 11 - Distribuição das solicitações por categorias 94
Tabela 12 - Lista de substituições 96
Tabela 13 - Treino do modelo (k-NN variação dos valores de k) 97

Acrónimos

API: *Application Program Interface*

ARN: *Artificial Neural Network*

BD: *Base de Dados*

CI: *Configuration Items*

CMBD*: Configuration Manager Database*

CMM: Capability Maturity Model

CMMI*: Capability Maturity Model Integration for Services*

COBIT: *Control Objects for Information and related Technology*

DF: *Document Frequency*

DW: *Data Warehouses*

EM: *Expectation–maximization*

GSTI: *Gestão de Serviços de TI*

HCI: *Human-Computer Interaction*

IDF: *Inverse Document Frequency*

IE: *Information Extraction*

IR: *Information Retrieval*

ITIL: *Information Technology Infrastructure Library*

ITSCMM: *Information Technology Services Capability Maturity Model*

KDT: *Knowledge Discovery from Texts*

KPI: *Key Performance Indicator*

OLAP: *Online Analytical Processing*

PMF: *Process Maturity Framework*

POS: *Part of Speech*

SDP: *Service Design Package*

SLA: *Service Level Agreement*

SLR: *Service Level Requirements*

SOM: *Self-Organizing Maps*

SVM: *Support Vector Machines*

TF: *Term Frequency*

TI: *Tecnologias de Informação*

VSM: *Vector Space Model*

WSD: *Word Sense Disambiguation*

1 Introdução

1.1 Contexto / Problema

A crescente utilização de recursos suportados por TI (Tecnologias de Informação), aliado à explosão de conteúdos e informação que surge a cada minuto, investe de responsabilidades acrescidas as estruturas que asseguram o seu suporte, gestão e manutenção. Desta forma, e para que se torne possível acompanhar de forma eficaz esta crescente utilização de recursos é imprescindível garantir que o suporte técnico seja adequado à utilização dos equipamentos e serviços, assegurando a fiabilidade, segurança e eficácia desejáveis.

Paralelamente, o aumento exponencial de dispositivos que se ligam nos dias de hoje às redes informáticas corporativas e que permitem aceder à informação de novas e variadas formas, obriga a um acrescido esforço na implementação de requisitos de segurança e mobilidade imprescindíveis à sua utilização, garantindo por um lado a segurança dos equipamentos e informação, e por outro a necessária flexibilidade na utilização das tecnologias.

Assiste-se diariamente contudo à necessidade de redução de recursos, quer tecnológicos, quer humanos para promover o suporte destas realidades, pela via financeira ou de contenção de mão-de-obra, o que parece contrariar a tendência natural da evolução do estado da arte.

É então óbvio que, se por um lado se obtém uma maior utilização dos recursos, por outro, se assiste a uma diversificação das formas de acesso, aliado à redução de meios de suporte e humanos, é crescente a necessidade de suporte informático para esta nova realidade, o que consequentemente obriga a repensar a estratégia de TI e consequentemente analisar e otimizar os seus processos.

O que não se pode medir, não se pode gerir (Drucker P. , 1999), e consequentemente o que não se pode gerir não se pode melhorar.

Um dos problemas identificados nesta temática é a dificuldade ou ausência de medição e registo nos processos de suporte a incidentes (disrupções ou potenciais falhas na disponibilidade ou qualidade dos serviços prestados) de TI, tornando-se a identificação dos pontos a melhorar mais complexa já que não são conhecidas com exatidão as proporções da categorias de incidentes, o que não permite por exemplo uma abordagem de com 20% de esforço resolver 80% das situações.

Aliado a esta escassez de registo de incidentes, a dispersão ou ausência de canais formais para assegurar centralizadamente esta gestão leva a situações de ausência de esquecimento, de resoluções divergentes para situações idênticas, de falhas de segurança e controlo e a uma diminuição global da perceção do trabalho realizado.

Outra das dificuldades prende-se com uma ausência na definição clara e objetiva de prioridades na resolução dos incidentes, sendo organizadas muitas vezes caso a caso de acordo com a posição hierárquica do solicitante ou do seu grau de envolvimento pessoal com o técnico ou gestor, o que restringe a automatização dos processos e consequentemente a sua melhoria global.

1.2 Questão de investigação

Observando o contexto e problemáticas identificados à luz da inteligência semântica, surge a questão se existe lugar à otimização destes processos, e caso exista, de que forma pode ser melhorada. Assim, a presente dissertação pretende encontrar a resposta para a seguinte questão:

É possível melhorar a eficácia e eficiência da gestão de uma implementação ITIL [1] com recurso à inteligência semântica?

1.3 Objetivos da investigação

Para a obtenção de possíveis respostas para a questão de investigação definida, importa traçar os objetivos gerais e específicos desta dissertação, que visam comprovar a hipótese colocada.

Identificam-se de seguida o objetivo geral e os específicos.

[1] ITIL – Information Technology Infrastructure Library

Objetivo geral

Definir e apresentar a aplicação de um modelo de inteligência semântica para apoio aos processos de gestão de serviços TI.

Objetivos específicos

a) Identificar e classificar o perfil dos serviços de TI relevantes;
b) Analisar o estado da arte relativamente aos modelos de inteligência semântica aplicáveis;
c) Modelar os processos de inteligência semântica;
d) Validar o modelo numa implementação ITIL.

1.4 Metodologia de estudo e pesquisa

O Estudo de Caso (Martins & Belfo, 2009) foi a metodologia de investigação escolhida para esta tese para aferição da qualidade do modelo a conceber, recorrendo-se também, de forma menos intensiva à metodologia "*Action Research*" (Martins & Belfo, 2009) (Baskerville, 1999) para criação, inferência e teste do modelo a conceber.

O estudo de caso ("*Case Study*") mostra-se potencialmente valioso entre os diversos métodos qualitativos de investigação, especialmente nas problemáticas que envolvem a utilização das tecnologias de contexto das organizações.

Não existindo uma definição clara e unanimemente aceite pela comunidade científica de estudo de caso, tal aproximação pode-se inferir (Fagundes, 2010) pela listagem das principais características, suas forças e fraquezas. Assim, conclui-se que o estudo de caso é um método de investigação que examina um fenómeno social no seu ambiente natural, através da recolha e análise de material empírico a partir de locais sociais específicos (ex: organizações), tendo como objetivos fundamentais, o alargar ou aprofundar o conhecimento científico sobre determinados fenómenos sociais, o poder construir uma teoria ou testar conceitos teóricos e relações entre os mesmos.

Na tabela 1 destacam-se as principais características do método de estudo de caso.

Tabela 1 - Características principais do método estudo de caso

O fenómeno é examinado no seu ambiente natural.
Os dados são recolhidos através de diversos meios.
Uma ou poucas entidades são examinadas (pessoa, grupo ou organização).
A complexidade da unidade é estudada intensivamente.
Os estudos de caso são mais aconselhados para a exploração, a classificação e nos diversos passos de desenvolvimento de hipóteses associados ao processo de construção do conhecimento; o pesquisador deve ter uma atitude recetiva para a exploração.
Não há envolvimento de nenhum controlo experimental ou manipulação.
O investigador poderá não especificar previamente o conjunto de variáveis independentes e dependentes.
Os resultados obtidos dependem muito do poder de integração do investigador.
Podem ocorrer mudanças na escolha do local e nos métodos de recolha de dados quando o investigador desenvolve novas hipóteses.
O estudo de caso é útil no estudo das questões "porquê" e "como" porque lidam com ligações operacionais para ser seguidas ao longo do tempo em vez de por frequência ou incidência.
O foco está nos acontecimentos atuais.

Fonte: *Adaptado de (Benbasat, Goldstein, & Mead, 1987)*

A partir dos finais dos anos 90, o método de investigação "*Action Research"* tornou-se popular para utilização em investigações académicas de sistemas de informação. Este método produz resultados muito relevantes, uma vez que é baseada em ações práticas cujo intuito é o de encontrar soluções imediatas para os problemas sustentadas na teoria em análise. Esta metodologia é composta por 5 fases (ilustração 1).

Ilustração 1 - Metodologia Action Research

Fonte: adaptado de (Baskerville, 1999)

A metodologia *Action Research* assenta nas seguintes fases (Baskerville, 1999):

- **Diagnóstico**: a fase de diagnóstico corresponde à identificação dos problemas primários que originam a necessidade da mudança.
- **Planeamento de ações:** nesta fase são identificadas as ações que podem atenuar ou resolver os problemas primários identificados na fase de diagnóstico. A descoberta e planeamento destas ações são sustentadas pela base teórica que definem o estado futuro desejado, bem como as ações que permitem atingir esse estado.
- **Tomada de ações:** é a fase de implementar as ações planeadas, efetivando as alterações propostas.
- **Avaliação:** nesta fase avaliam-se os efeitos teóricos das ações que foram realizadas, e se de facto contribuíram para a resolução ou atenuação dos problemas. Mesmo quando as alterações efetuadas resultam em sucesso, deve na

fase de avaliação questionar-se e verificar se os resultados se devem exclusivamente às ações tomadas, ou se existiram outros efeitos paralelos infligidos pelas próprias das ações que influenciaram os resultados.

- **Documentar a Aprendizagem:** embora a atividade de documentação da aprendizagem surja como a última das cinco fases, a recolha de dados e informação é um processo continuo que ocorre ao longo de toda a metodologia, consolidando-se nesta fase a informação recolhida em conhecimento. O conhecimento obtido pode ser dirigido para três audiências:
    - Redefinição de normas e procedimentos para utilização pela organização e processo;
    - Quando a alteração não teve sucesso, o conhecimento adicional obtido pode providenciar fundamentações futuras para diagnosticar e preparar novas intervenções de *action research*.
    - Finalmente, o sucesso ou insucesso das *frameworks* teóricas criadas providenciar conhecimento importante para a comunidade científica preparar e lidar com investigações futuras.

De acordo com esta metodologia, será desenhado um modelo e escolhida uma organização onde, recorrendo às práticas de ITIL implementadas, a maturidade do modelo poderá ser aferida e o modelo estudado. Após esta avaliação, os resultados serão estudados e obtidas conclusões e inferidas melhorias no modelo com base nas mesmas.

A opção pela metodologia de investigação de Estudo de Caso aliada à técnica de *Action Research* deve-se sobretudo, ao facto de tornar possível a conceção e otimização de um modelo, que será posteriormente aplicado e testado num caso concreto.

Considera-se assim que para a estratégia ter sucesso é essencial o foco em fenómenos contemporâneos inseridos em algum contexto da vida real. Aspetos importantes como questões de pesquisa; definição da amostra; dados para recolha e análise dos resultados deverão ser considerados para aumentar a validade do Estudo de Caso.

1.5 Organização da dissertação

O presente estudo integra diferentes dimensões da análise semântica, transportando-as para a aplicação real em problemas tangíveis e concretos à luz de uma *framework* reconhecida como é a ITIL. Assim, assiste-se a uma primeira fase de contextualização dos problemas a resolver, a descrição da metodologia a utilizar, passando-se à exposição dos conceitos teóricos base de sustentação às ações que visam resolver ou atenuar os problemas identificados. Com base nestas ações e na metodologia escolhida, será identificado o modelo conceptual e aferida a sua maturidade e aplicabilidade a um estudo de caso. Para cobrir os temas referidos, optou-se pela seguinte estrutura:

Capítulo 1 - Introdução

Este capítulo tem por objetivo alinhar o *focus* no contexto e envolvente da tese, abordando o problema a analisar e endereçar e vertendo-o em hipóteses de investigação, que serão posteriormente investigadas e documentadas recorrendo às respetivas metodologias de estudo. No final deste capítulo, é detalhada a organização da tese descrevendo-se o conteúdo por cada capítulo.

Capítulo 2 - Revisão de Literatura

O capítulo 2 pretende destacar da literatura de referência na área os pontos-chave mais relevantes de forma sucinta e organizada, permitindo aculturar as ideias e conceitos imprescindíveis às restantes fases da metodologia. Inicia-se por uma revisão de conceitos e informação relativa à ITIL, abordando-se o tema da gestão da mudança e finalmente a análise de conteúdos semânticos incluindo a abordagem dos conceitos inerentes aos diversos ramos desta ciência aplicáveis a esta tese.

Capítulo 3 - Metodologia de Investigação

Após identificação dos objetivos e revista a literatura de referência segue-se o capitulo 3, onde será descrita com maior detalhe a metodologia de investigação pretendida, bem como as respetivas estratégias e métodos.

Capítulo 4 - Desenvolvimento do Modelo de Inteligência Semântica

Será neste capítulo que se irá dar forma ao modelo de inteligência semântica, assente nas metodologias de investigação anteriormente identificadas. Pretende-se que o modelo resultante seja implementável e abrangente por forma a ser aplicável ao maior número de casos possível.

Capítulo 5 - Estudo de Caso

Concebido e desenhado o modelo, testa-se neste capítulo a sua aplicabilidade e utilidade, aplicando-o a estudo de caso concreto, e obtendo dessa aplicação resultados e métricas que permitam inferir a performance do modelo e obter algumas conclusões relativas ao modelo e sua aplicação.

Capítulo 6 - Conclusões e Trabalhos Futuros

Neste capítulo serão apresentadas os principais resultados e conclusões deste estudo, bem como elencados alguns pontos de trabalho futuro que, ou por não estarem no âmbito desta dissertação, ou pelos motivos identificados, não foram alvo de análise, remetendo-se portanto para trabalhos futuros.

2 Revisão de Literatura

2.1 ITIL (*Information Technology Infrastructure Library*)

As organizações em todo o mundo estão cada vez mais dependentes na fiabilidade dos seus processos de negócio. Pequenas falhas numa pequena componente de um sistema poderão resultar no mau funcionamento de componentes cruciais nos processos de negócio (Abramowicz, Witold et all, 2007) e consequentemente originar grandes perdas. Desta forma, como podem o negócio e as equipas de TI providenciar respostas céleres e estabilidade aos sistemas que suportam o negócio? Quem será diretamente afetado? Qual é o impacto de uma falha nos processos de negócio? Qual a melhor forma de restabelecer o sistema? Diversos autores sugerem a adoção de processos de gestão baseados nas boas práticas de ITIL (Abramowicz, Witold et all, 2007). Esta *framework* de boas práticas disponibiliza uma base estável de classificação e descrição de sistemas, assente em ***Configuration Items*** (CIs: módulos de *hardware*, *software* ou recursos humanos), o que facilita a descoberta, especificação, implementação, controlo e monitorização dos processos (Abramowicz, Witold et all, 2007) (Ward & Peppard, 2002).

A ITIL não é inquestionável, até porque a quantidade de práticas diferentes implementadas pelo mercado mostra que é possível conseguir bons resultados aplicando outro tipo de normas e de processos.

Ao implementar a ITIL é necessário ter em consideração, por um lado os benefícios gerados no nível de operacionalidade, e por outro o posicionamento estratégico obtido perante os clientes. Ter uma opinião informada sobre as melhorias que podem ser produzidas por estas *frameworks* é importante para os investigadores e utilizadores (Marrone & Kolbe , 2011).

Neste estudo (Marrone & Kolbe , 2011), através da aplicação de *frameworks* de avaliação de ITSM (*IT Service Management*), é demonstrado que os benefícios conseguidos são proporcionais ao nível de maturidade da implementação de ITIL, bem

como às métricas de medição dos próprios benefícios, garantindo paralelamente uma melhor perceção de alinhamento entre o negócio e as TI.

2.1.1 Histórico

No final dos anos 1980 a *Central Computer and Telecommunications Agency* (CCTA), atualmente OGC (*Office for Government Commerce*), órgão do governo da Inglaterra, recolheu e analisou informações de diversas organizações e selecionou as orientações mais úteis para a CCTA e seus clientes no governo britânico, tendo por base a necessidade do governo de ter processos organizados na área de TI (Axelos, 2015).

Deste estudo resultou um livro de orientações para ser aplicado em empresas relacionadas com o governo. Rapidamente as empresas não relacionadas com o governo entenderam que essas orientações eram genéricas e também aplicáveis aos seus negócios e ambientes, passando a adotá-las. O resultado deste processo foi a compilação de uma biblioteca de melhores processos e melhores práticas de prestação e gestão de serviços de TI, que passou a ser conhecido como a ITIL – *Information Technology Infrastructure Library*.

A ITIL seguiu sua versão original, composta por 31 volumes. No início dos anos 2000, foi publicada a versão 2 da ITIL, formada por um conjunto de 7 volumes que já incluem uma visão global sobre boas práticas para prestação de serviços de TI, tornando-se numa base padrão para a norma BS 15000, que se tornou um anexo da norma ISO 20000. No início de 2007, foi publicada a versão 3 da ITIL, também conhecida como *ITIL Refresh Project*, formada por apenas cinco volumes que compilaram os pontos fortes nas versões anteriores, organizados sobre conceitos relativos a uma estrutura de ciclo de vida de serviços. Em Julho de 2013 a ITIL foi adquirida pela AXELOS, Lda, entidade que é agora responsável pela atribuição de licenças de utilização da propriedade intelectual da ITIL, sistemas de acreditação e exames e atualizações à mesma.

2.1.2 Conceitos e definições

A ITIL disponibiliza uma biblioteca comum para a gestão dos serviços de TI de uma organização, descrevendo melhores práticas de forma coesa. Os livros abordam como

estas práticas podem ser otimizadas e como a coordenação das atividades pode ser aperfeiçoada, descrevendo e explicando também como os processos podem ser formalizados dentro de uma organização, padronizando uma terminologia comum para os serviços de TI e ajudando a definir os objetivos e determinar o esforço requerido.

As melhores práticas da ITIL têm como objetivos:

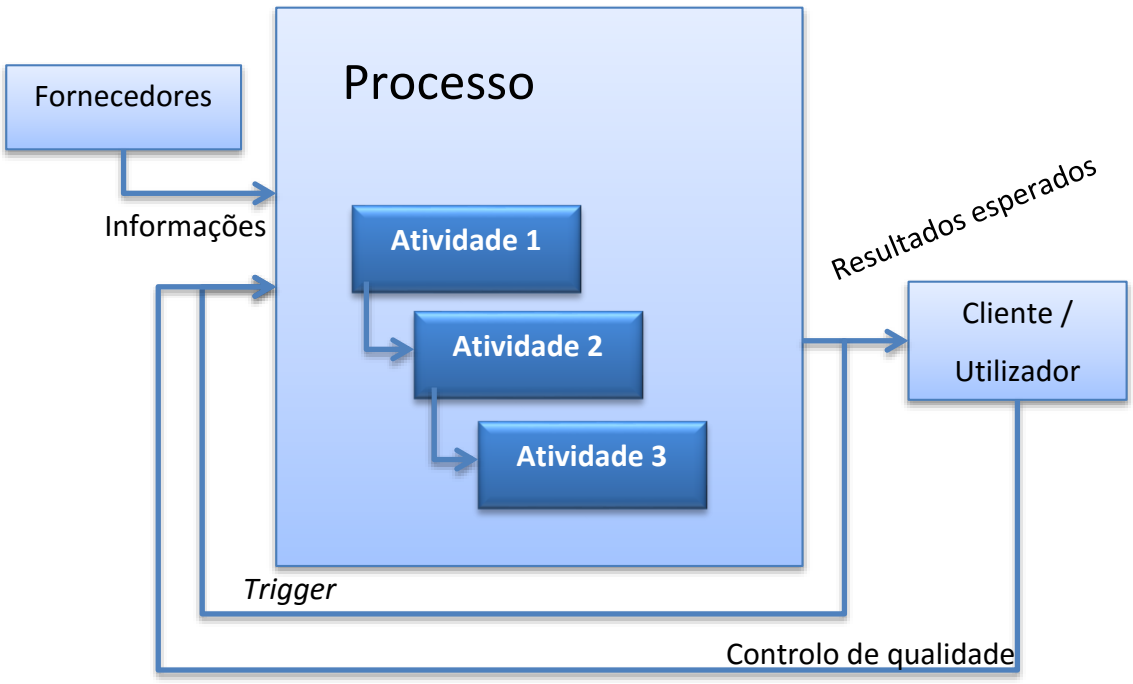
- Reduzir custos;
- Aumentar a disponibilidade;
- Ajustar a capacidade;
- Aumentar a eficiência e eficácia;
- Melhorar a escalabilidade;
- Reduzir riscos.

A ITIL não pretende ser vista como uma metodologia, pois as melhores práticas são flexíveis a ponto de poderem ser adaptadas a quaisquer processos (algumas mesmo a processos que não são de TI); já uma metodologia pressupõe uma implementação mais rígida, com regras bem definidas.

Neste contexto, (Axelos, 2012) entende-se **processo** enquanto um conjunto de atividades inter-relacionadas com objetivos definidos. Possui entradas de dados, informações e produtos para, através da identificação dos recursos necessários ao processo, transformar estas entradas nos objetivos previstos. Hoving e van Bon (W. & van Bon, 2008) acrescentam que, um processo descreve o que uma organização faz. O glossário do ITIL v3 (Crown, 2011) define processo como, um conjunto estruturado de atividades destinadas a atingir um objetivo específico. Um processo contém uma ou mais entradas definidas e transforma-as em saídas definidas. Pode incluir cargos, responsabilidades e controlos de gestão necessários para fielmente dirigir as saídas. Um processo pode definir políticas, normas, orientações, atividades e instruções de trabalho. Contudo, outra publicação do ITIL, o volume ITIL v3 *Service Strategy* (Axelos, 2015), descreve o termo processo de forma ligeiramente diferente: "*é um conjunto de atividades coordenadas combinando e executando recursos e capacidades, a fim de*

produzir um resultado que, direta ou indiretamente, cria valor para um cliente externo ou um grupo envolvido.", como descrito na ilustração 2.

Ilustração 2- Funcionamento de um processo básico segundo o ITIL v3



Fonte: Adaptado de (W. & van Bon, 2008)

Hoving & van Bon (W. & van Bon, 2008) declaram que, "*A gestão de processos é geralmente compreendida como a obtenção do caminho mais curto para alcançar o valor dos clientes.*"

Paralelamente, o foco da ITIL assenta em serviços e sua gestão eficiente e melhoria continua, com vista à satisfação do cliente / utilizador. Os **serviços** apresentam-se de maneira diferente aos clientes e utilizadores, comparativamente aos produtos, e podem ser decisivos pela sua interatividade e proximidade no momento de efetivar um contrato. Contudo, produtos e serviços aparecem relacionados inúmeras vezes, numa dependência recíproca. As organizações destacam-se cada vez mais pelos serviços prestados aos clientes, ao seu público-alvo ou mesmo aos seus próprios utilizadores. O itSMF (itSMF, 2007) reforça esta ideia ao afirmar que um serviço é um meio de entregar valor aos clientes, facilitando os resultados que os clientes querem alcançar,

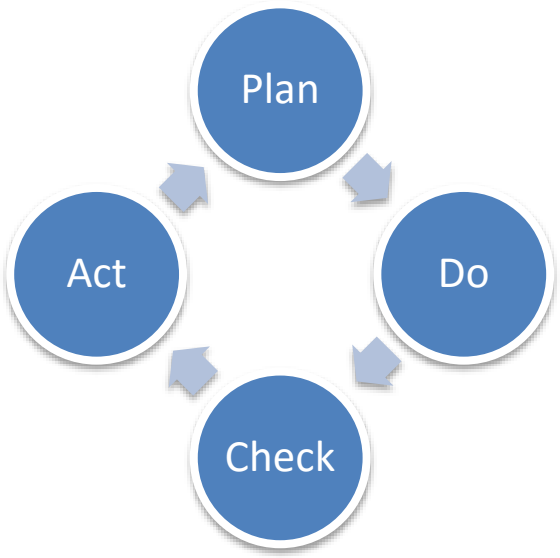
sem ter que assumir riscos e custos. Quanto às suas caraterísticas, os serviços podem divergir muito, mas alguns atributos são vulgarmente referenciados, tais como:

- Impossibilidade de armazenar;
- Os clientes usualmente estão envolvidos, de alguma forma, na sua produção;
- Intangibilidade;
- Heterogeneidade;
- Suscetíveis de perecer;
- Difícil medir a sua qualidade.

A biblioteca ITIL na versão 3 aproxima-se à **Gestão de Serviços TI** (GSTI) e o seu glossário (Crown, 2011) define GSTI como "*A implementação e a gestão da qualidade dos serviços de TI de forma a atender às necessidades de negócio. A gestão de serviços TI é feita pelos fornecedores de serviço de TI por meio da combinação adequada de pessoas, processos e tecnologias da informação.*". Fagundes (Fagundes, 2010), resume muito bem esta disciplina ao afirmar que "*A gestão de serviços de TIC é um conjunto de disciplinas que oferece o serviço certo a um custo certo, dentro de níveis de qualidade e prazos que vão de encontro às expectativas dos negócios*" e acrescenta ainda "*Gestão de Serviços é uma área emergente e pouco compreendida por muitas pessoas. Os líderes de TI sentem a pressão para melhorar significativamente o desempenho do serviço, mas enfrentam uma quantidade surpreendente de informações contraditórias e incompletas*".

Na ITIL a abordagem aos serviços está intrinsecamente ligada ao ciclo de Deming (Deming, 1986) nas suas quatros fases: *Plan* (planear), *Do* (executar), *Check/Control* (verificar) e *Act* (agir), encontrando-se as duas últimas fases orientadas à melhoria contínua do modelo.

Ilustração 3 - Ciclo de Deming (PDCA)



Fonte: adaptado de (Deming, 1986)

Apresentando as fases da ilustração 3 em mais detalhe, temos:

- **Planear (*Plan*)**: definir os objetivos e processos necessários para entregar resultados de acordo com as necessidades dos clientes e das políticas da organização;
- **Executar (*Do*):** implementar o plano, executar o processo, fazer o produto. Nesta fase recolhem-se dados e indicadores para mapeamento e análise nas fases Controlar e Agir;
- **Controlar (*Check*):** Analisar os resultados obtidos e medidos na fase anterior, compará-los com os resultados esperados (e definidos na primeira fase) e determinar as diferenças e desvios.
- **Agir (*Act*)**: Tendo presente os resultados da comparação da fase anterior, torna-se possível determinar as causas das diferenças e definir mudanças ou melhorias no processo ou produto, quer através de ações corretivas ou de melhoria.

Na terminologia ITIL (ITIL Service Management, 2013), um ***Service Level Agreement*** (SLA) (Acordo de Nível de Serviço) é um acordo entre um fornecedor de um serviço e o seu cliente, que descreve o serviço em causa, as suas metas, os papéis e responsabilidades dos intervenientes e partes envolvidas no acordo. Este acordo tem por

base os requisitos de serviço, ***Service Level Requirements*** (SLR) que definem, recorrendo a indicadores mensuráveis e ***Key Performance Indicator*** (KPI) os valores mínimos a cumprir.

Ainda no âmbito da ITIL, um **incidente** define-se por uma interrupção não planeada de um serviço de TI, a redução da qualidade de um serviço de TI tendo por base os SLA acordados, ou a falha de um CI que não tenha ainda tido impacto num serviço de TI.

Por exemplo, falha de um disco rígido, mesmo quando existe um sistema de redundância, é considerada um incidente. Um **problema** é uma falha cuja causa é conhecida mas para a qual não existe ainda uma solução.

Um **pedido** refere-se a qualquer alteração ao ambiente ou recursos de TI, seja a atribuição de equipamento, alteração de *password*, etc.

2.1.3 Volumes do ITIL

Tal como referido anteriormente, o ITIL na versão 3 encontra-se organizado em volumes temáticos, cada um endereçando objetivos específicos e suportado em processos característicos, que condensam e otimizam os pontos fortes das versões anteriores, sendo os seguintes:

- *Service Strategy* (Estratégia do Serviço)
- *Service Design* (Desenho de Serviço ou Projeto de Serviço)
- *Service Transition* (Transição do Serviço)
- *Service Operation* (Operação do Serviço)
- *Continual Service Improvement* (Melhoria Contínua do serviço)

Nos pontos seguintes será caracterizado em maior detalhe cada um dos livros.

2.1.3.1 *Service Strategy (Estratégia de Serviço)*

A Estratégia de Serviço resulta da aplicação efetiva da estratégia e cultura da organização. Tem por base o conceito da qualidade de serviço desejada pelos clientes, ao invés da simples disponibilização de um produto, para satisfazer necessidades específicas. Assim, o serviço prestado pela organização ou entidade deverá focar-se em

garantir o valor acrescentado do serviço que é efetivamente percecionado pelo cliente alvo, partindo do pressuposto de que se conhece de forma profunda o mercado ou negócio e os clientes para que se opera.

A estratégia de serviço deverá ser o pilar central do ciclo de vida da gestão de serviços de TI, sendo usado como plano base para as restantes fases, definindo as melhores alternativas de solução.

O resultado da estratégia de serviço deverá passar por um conhecimento profundo de que serviços são oferecidos, a quem e de que forma o devem ser, sabendo exatamente aquilo que faz os clientes escolherem os serviços da própria organização em detrimento de outras.

Deverão ainda ser claros os custos provenientes da implementação da estratégia definida, pela alocação dos recursos necessários ao cumprimento e otimização do conjunto de serviços prestados. É também nesta fase que se deverá conceber o modelo de avaliação e desempenho do serviço e respetivas métricas, para que seja possível a sua medição e eventual posterior melhoria.

Por fim, e concebida a forma para o reconhecimento do valor percecionado pelos clientes, importa que esta estratégia seja também acolhida, apoiada e valorizada pelos gestores da organização, reconhecendo-lhe as vantagens e apostando nela como uma oportunidade de negócio, dando lugar a investimentos seguros na oferta de serviços e na capacidade de os gerir.

Neste volume da ITIL são definidos alguns conceitos chave da filosofia da biblioteca, dos quais se destacam os quatro P's da estratégia: perspetiva, posicionamento, plano e padrão.

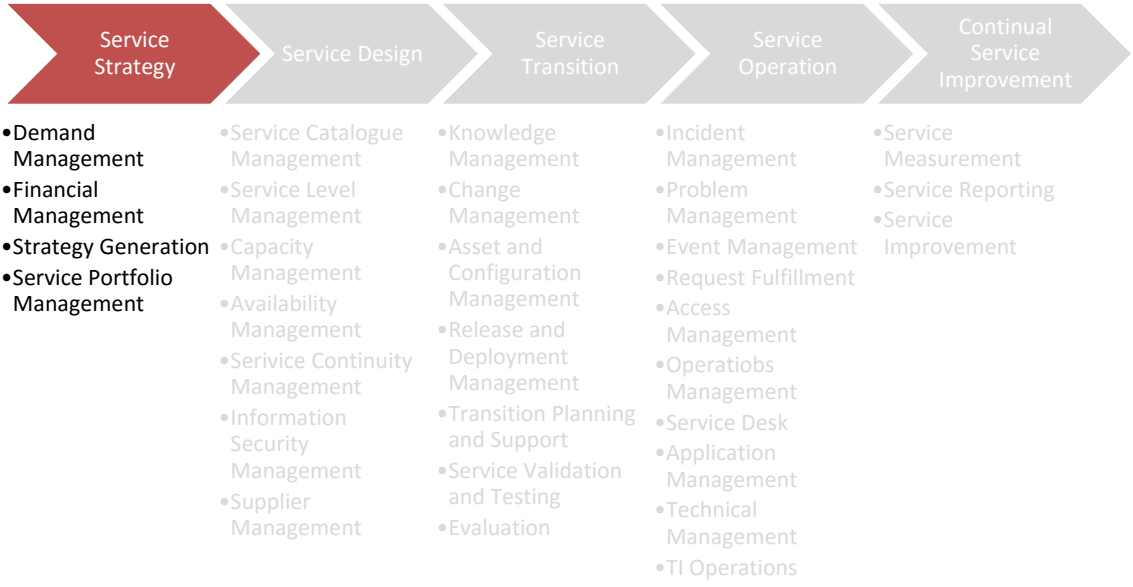
A este conceito juntam-se alguns como os de competição e mercado, qualidade e valor de serviço, definição de fornecedores e clientes, bem como a identificação do *Service Management* como ativo estratégico das organizações.

Este livro inclui os processos:

- *Demand Management*
- *Financial Management*
- *Strategy Generation*
- *Service Portfolio Management*

Nesta fase têm intervenção o Gestor de Relações de Negócio, o Gestor de Produto e o Diretor de Aquisições.

Ilustração 4 - Livros do ITIL: Service Strategy



Fonte própria baseado em (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007)

2.1.3.2 *Service Design (Desenho de Serviço)*

O Desenho de Serviço é uma das fases necessariamente transversal a todo o ciclo de vida do serviço, tendo por objetivo conceber "*o desenho de serviços de TI apropriados e inovadores, desde a arquitetura, processos e regras, até à documentação, com o objetivo de agrupar os requisitos de negócio acuais e de futuro*" (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007).

O objetivo associado a esta fase passa pela identificação e gestão de possíveis riscos, sempre com vista à obtenção dos resultados de negócio esperados.

Esta fase determina a existência do desenho das infraestruturas, ambientes, aplicações e recursos necessários. Após essa análise estima-se que estejam reunidas as condições para definir os métodos e métricas de análise da implementação e manutenção de processos, regras, normas, arquiteturas e *frameworks*. Todas estas implementações deverão ser sustentadas por documentação.

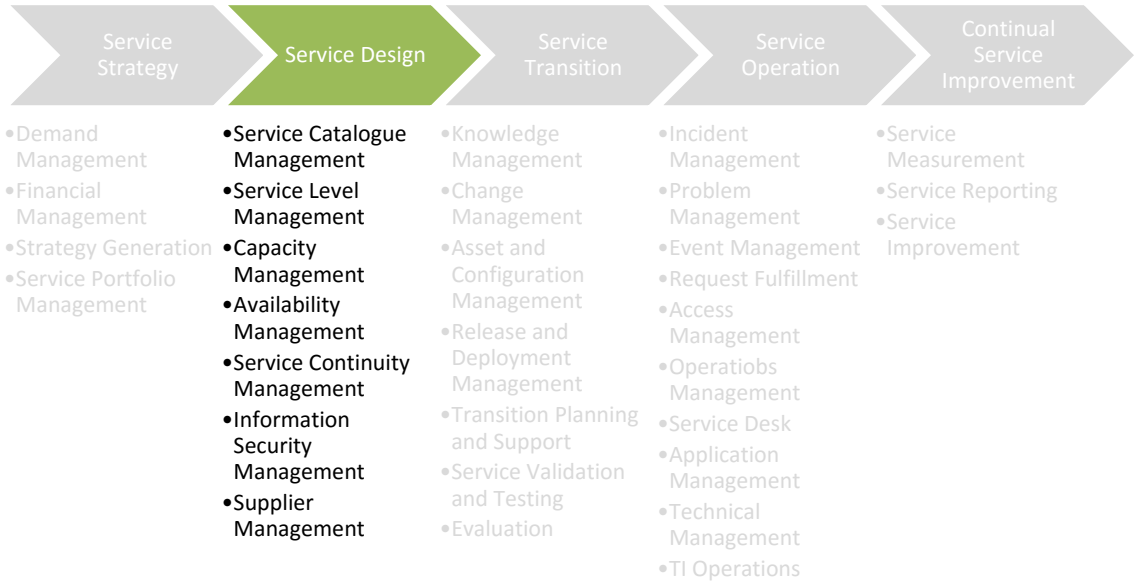
Com base nos requisitos identificados anteriormente, o desenho de serviços visa o desenvolvimento da criação de uma solução à medida, criando nesse processo o *Service Design Package* (SDP), que inclui todos os desenhos e documentos de suporte aos serviços e processos produzidos, e os seus requisitos ao longo de todas as fases do seu ciclo de vida.

Este livro inclui os processos:

- *Service Catalogue Management* (Gestão do Catálogo de Serviços)
- *Service Level Management* (Gestão de Níveis de Serviços)
- *Capacity Management* (Gestão de Capacidade)
- *Availability Management* (Gestão de Disponibilidade)
- *Service Continuity Management* (Gestão de Continuidade de Serviço)
- *Information Security Management* (Gestão da Segurança da Informação)
- *Supplier Management* (Gestão de Fornecedores)

Desta fase fazem parte o Gestor de Desenho de Serviço, Designer/Arquiteto de TI, Gestor de Catálogo de Serviços, Gestor de Nível de Serviço, Gestor de Disponibilidade, Gestor de Continuidade de Serviço de TI, Gestor de Capacidade, Gestor de Segurança e Gestor de Fornecimento.

Ilustração 5 - Livros do ITIL: Service Design



Fonte própria baseado em (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007)

2.1.3.3 *Service Transition (Transição do Serviço)*

Tendo por base os requisitos definidos e detalhados nas fases anteriores, principalmente o SDP, na fase Transição de Serviço é preparada a operacionalização dos serviços, colocando assim em prática o plano.

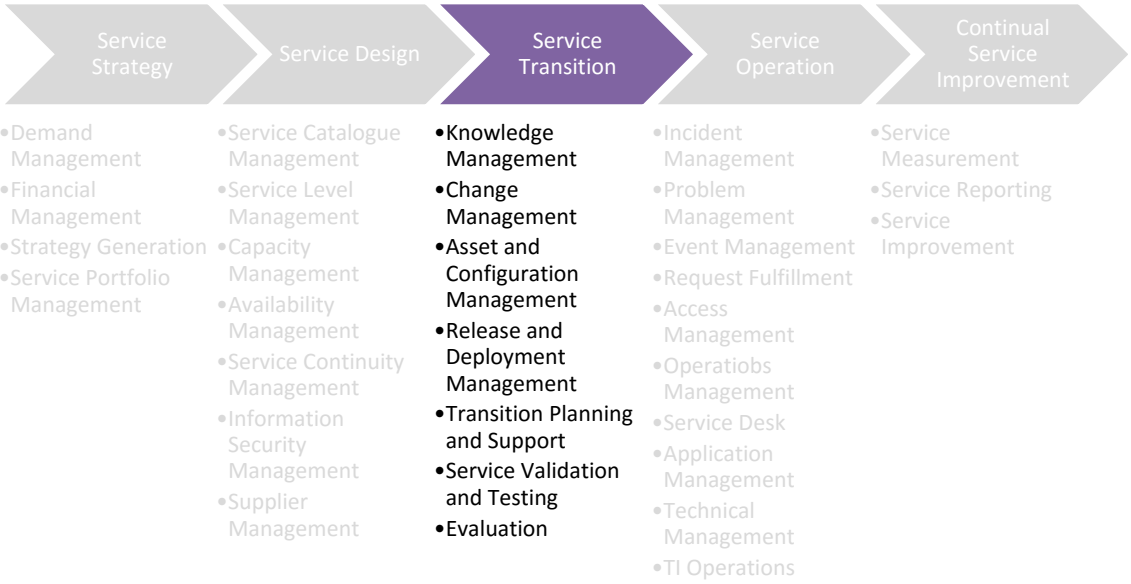
Esta fase pressupõe uma estabilidade no que concerne às regras de negócio, pelo que é crítica que a sua conclusão ocorra após garantia de que as necessidades da organização e clientes foram exaustivamente analisadas. As regras de introdução de alterações, deste passo para a frente, deverão também ser descritas na fase de transição.

Após este exigente processo, devem ser definidas as situações extremas, ou circunstâncias limite, em que esses serviços poderão operar, estabelecendo-se limites mínimos aceitáveis de recursos, explorando paralelamente possibilidades excecionais que possam ocorrer.

Este livro inclui os processos:

- *Knowledge Management*
- *Change Management*
- *Asset and Configuration Management*
- *Release and Deployment Management*
- *Transition Planning and Support*
- *Service Validation and Testing*
- *Evaluation*

Ilustração 6 - Livros do ITIL: Service Transition



Fonte própria baseado em (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007)

2.1.3.4 *Service Operation (Operação do Serviço)*

A fase de *Service Operation* é a última fase que complementa a componente funcional de implementação de serviços. Trata, após a transição, da operacionalização dos processos identificados como necessários para a prestação dos serviços requisitados pelas regras de negócio da organização.

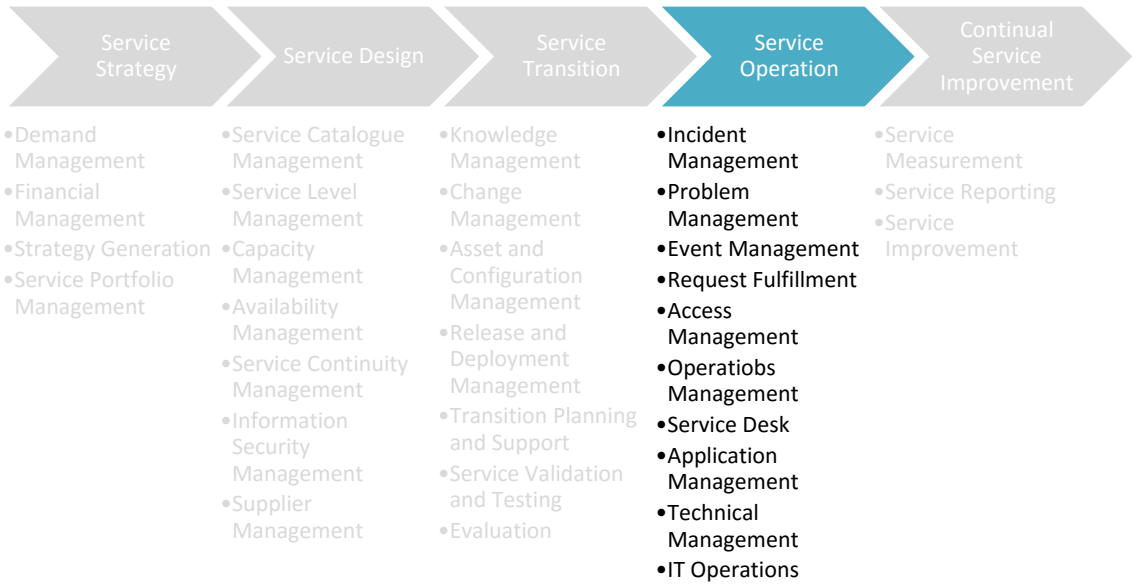
Esta fase tem por objetivos finais a implementação dos serviços e qualidades acordadas para os mesmos, assim como a gestão de aplicações, infraestruturas e tecnologia de base aos mesmos.

Uma das dificuldades nesta implementação passa pelas decisões a serem tomadas pelos gestores, que, segundo este volume do ITIL, devem envolver quatro *tradeoffs* essenciais: TI interno ou Negócio externo, estabilidade ou capacidade de resposta, qualidade de serviço ou custo do mesmo, reação ou pro-atividade. Tendo por base estas decisões, são implementados, operacionalizados e prestados os serviços acordados.

Este livro inclui os processos:

- *Incident Management* (Gestão de Incidentes)
- *Problem Management* (Gestão de Problemas)
- *Event Management* (Gestão de Eventos)
- *Request Fulfillment* (Gestão de Pedidos)
- *Access Management* (Gestão de Acessos)
- *Operations Management* (Gestão de Operações)
- *Service Desk*
- *Application Management* (Gestão de Aplicações)
- *Technical Management* (Gestão de Técnica)
- *TI Operations* (Operações de TI)

Ilustração 7 - Livros do ITIL: Service Operation



Fonte própria baseado em (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007)

2.1.3.5 *Continual Service Improvement (Melhoria Continua do Serviço)*

Concluídas as fases de operacionalização dos serviços, cabe à organização seguir a filosofia inerente ao ITIL e estruturar um processo de melhoria contínua de serviços, completando assim todas as fases do ciclo de Deming.

Percebendo o valor compreendido pelos gestores e clientes da organização, no que concerne aos serviços prestados, cabe à organização manter os níveis de serviço pré-definidos. Assim, e compreendendo o mercado aberto e global, capaz de encontrar melhores serviços e soluções a que o cliente pode atribuir maior valor, é necessário que de forma contínua, e seguindo os melhores padrões existentes, a organização procure melhorar de forma contínua a qualidade do serviço prestado, evoluindo todo o sistema de serviços de TI e processos decorrentes do mesmo.

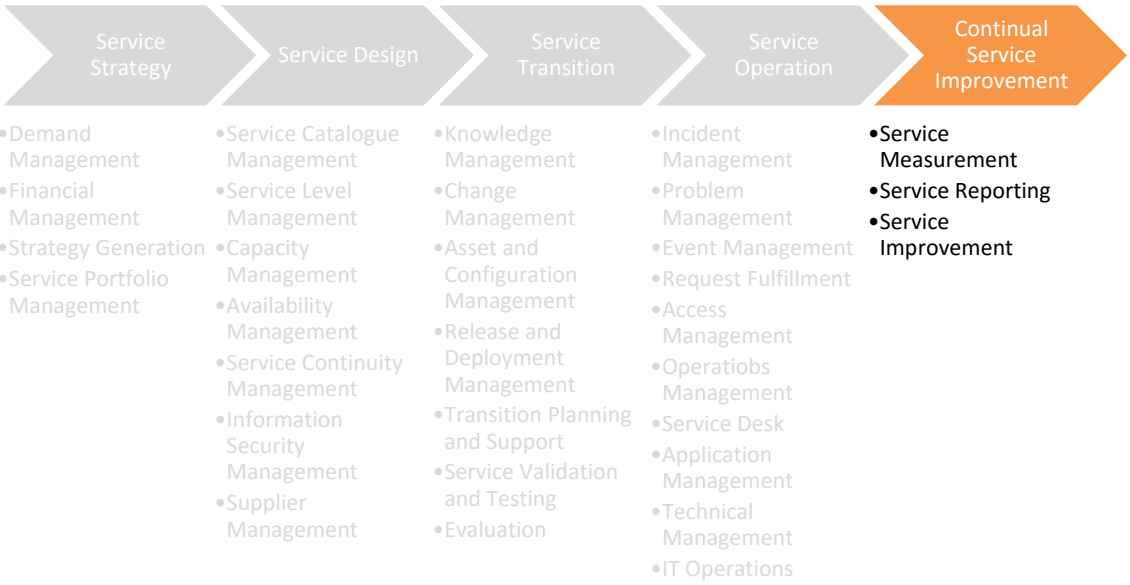
Este volume do ITIL é uma combinação dos princípios, práticas e métodos da gestão de qualidade, gestão de mudança e melhoria das capacidades, ao longo de todo o ciclo de vida do produto.

Nesta fase de maturidade dos serviços, é identificado um processo chave do *Continual Service Improvement*: os sete passos do processo de melhoria. Esses passos são descritos neste livro como a definição do que se deve medir, o que se pode medir, recolha da informação, processamento da mesma, posterior análise seguida de apresentação e utilização da mesma, tendo por último passo a implementação de ações corretivas de melhoria.

Este livro inclui os processos:

- *Service Measurement*
- *Service Reporting*
- *Service Improvement*

Ilustração 8 - Livros do ITIL: Continual Service Improvement



Fonte própria baseado em (Cartlidge, Hanna, Rudd, Macfarlane, Windebank, & Rance, 2007)

2.1.4 Modelos de maturidade

Neste ponto serão descritos alguns modelos de aferição de maturidade que visam definir uma escala de avaliação uniforme para o grau de maturidade da aplicação da ITIL

dentro de uma organização. Esta aferição permitirá não só, entre diferentes entidades, comparar o seus níveis de maturidade na aplicação da ITIL mas também, e no âmbito deste documento garantir a definição de um nível de maturidade mínima, para o qual o modelo a propor funcionará, criando assim uma *baseline* de medidas da ITIL implementadas e/ou recomendadas.

2.1.4.1 Capability Maturity Model (CMM)

Um dos primeiros modelos analisados é o Modelo de Capacidade e Maturidade, *Capability Maturity Model* (CMM). Este modelo foi desenvolvido pelo *Software Engineering Institute* (SEI) da universidade de Carnegie Mellon. Identifica cinco níveis de maturidade para avaliar um processo de uma organização através da aferição do grau em que o mesmo está definido e é gerido:

- **Nível 1: Inicial**. No nível inicial, a organização normalmente não tem um ambiente de TI estável, denominado muitas vezes *ad-hoc* ou caótico, sendo a capacidade de fazer determinada pelos elementos que compõem as equipas e não pela própria organização.
- **Nível 2: Repetível:** estão definidas políticas para gestão das TI e existem procedimentos definidos para as implementar. O planeamento e gestão de novos projetos são baseados no histórico de processos semelhantes.
- **Nível 3: Definido**: Os processos típicos e mais frequentemente usados encontram-se documentados, estando integrados como um todo. Neste nível as organizações podem-se considerar organizadas e consistentes, uma vez que as atividades estão estabilizadas e são repetíveis.
- **Nível 4: Gerido**. A organização define quantitativamente um conjunto de objetivos para os serviços e processos, e executa os processos aferindo métricas consistentes e bem definidas. Estas métricas estabelecem uma base quantitativa para a avaliação dos produtos e serviços.
- **Nível 5: Otimização**. Toda a organização está focada no processo de melhoria contínua, tendo a meios e capacidades para identificar as fraquezas e forças para melhorar proactivamente o processo, evitando assim falhas. A organização analisa as falhas e determina as causas, com vista à resolução da sua origem.

Cada nível de maturidade é identificado por um conjunto de itens que são classificados e mapeados em objetivos e recursos comuns a diversas organizações. Com base nesses itens e respetivos níveis, são aferidos os processos chave, identificado o nível de maturidade da organização. Para atender os objetivos, são colocadas em prática medidas para continuamente melhorar a qualidade na direção do nível de maturidade seguinte. Muito embora este modelo de aferição não seja totalmente adequado à gestão de ambientes de TI, âmbito deste documento, a sua aplicabilidade permite de forma transversal identificar um nível nas áreas relacionadas com o apoio a aplicações, já que este modelo trata em particular dos ciclos de desenvolvimento e maturidade do *software*, o que muitas vezes cruza com a gestão das TI.

2.1.4.2 Process Maturity Framework

A PMF (Framework de Maturidade de Processos ou *Process Maturity Framework)*, é descrita no livro da ITIL na sua v2, muito embora na revisão v3 não existam referências. Pode ser utilizada para aferir a maturidade de um serviço específico, ou a maturidade do *Service Management* como um todo (Coelho, 2009).

- Esta *framework* define cinco níveis como descrito na tabela 2, sendo cada nível caracterizado por uma combinação dos cinco elementos seguintes:
- Visão e rumo
- Processo
- Pessoas / Recursos Humanos
- Tecnologia
- Cultura

Tabela 2 - Níveis PMF e descrição

Nível	PMF	Foco	Comentários
1	Inicial	Tecnologia	Tecnologia, especialistas, excelência das soluções
2	Repetível	Produto, serviço	Processos operacionais, como o suporte
3	Definido	Cliente	Gestão eficaz de níveis de serviço
4	Gerido	Negócio	Estratégia de TI alinhada com o negócio
5	Otimizando	Cadeia de Valor	Integração perfeita entre o TI e a estratégia de negócio

Fonte própria

Cada nível tem um ou mais objetivos que são necessários implementar para que a organização o possa atingir.

2.1.4.3 Modelo de Maturidade em Capacitação (CMMI) for Services

O Modelo de Maturidade em Capacitação - Integração - *Capability Maturity Model Integration for Services (CMMI)* funciona enquanto guia de aplicação das melhores práticas no fornecimento de serviços numa organização (Coelho, 2009). As melhores práticas propostas pelo modelo focam-se na qualidade dos serviços prestados do ponto de vista dos clientes e utilizadores finais, integrando conhecimento que é essencial à prestação do serviço.

O CMMI desenha os conceitos e práticas com base em diversos *standards* e modelos, incluindo:

- Information Technology Infrastructure Library (ITIL)
- ISO/IEC 20000: Information Technology - Service Management
- Control Objects for Information and related Technology (CobiT)
- Information Technology Services Capability Maturity Model (ITSCMM)

O CMMI dispõe de duas abordagens: analisar toda a organização pelas etapas, ou por processos. Existem 24 processos que são caracterizados por práticas e objetivos específicos, existindo contudo objetivos e praticas que são transversais a todos os processos (Coelho, 2009).

Tabela 3- Níveis de análise por Etapas ou por processos

Nível	Análise continua (processos) Níveis de Capacidade	Análise por etapas Níveis de Maturidade
0	Incompleto	(não aplicável)
1	Executado	Inicial
2	Gerido	Gerido
3	Definido	Definido
4	Quantificado e gerido	Quantificado e gerido
5	Otimizando	Otimizando

Fonte própria

A descrição dos níveis da análise por processos é a seguinte:

- **Nível 0: Incompleto**. Um ou mais objetivos específicos na área de processo não são satisfeitos, ou não existem objetivos genéricos definidos.
- **Nível 1: Executado:** os objetivos específicos de uma parte do processo são satisfeitos. Suporta e define as tarefas necessárias para providenciar serviços.
- **Nível 2: Gerido:** neste nível existe planeamento e a execução é feita de acordo com uma política; emprega recursos humanos com capacidades e recursos adequados para produzir *outputs* controlados; existe envolvência dos gestores e decisores; existe monitorização, controlo e revisão;
- **Nível 3: Definido**. Os *standards*, descrições do processo e procedimentos estão definidos com base num conjunto *standard* de processos que servem de guia para novos projetos organizacionais.
- **Nível 4: Quantificado e gerido**. O processo é controlado através da análise estatística e quantitativa. Estão definidos objetivos quantitativos para a qualidade e performance do processo, sendo estes indicadores definidos como indicadores de gestão do processo.
- **Nível 5: Otimizando**. Processo é continuamente melhorado com base na análise das causas das condicionantes ou proactivamente melhorado com base nos indicadores de performance dos processos em curso.

A representação por etapas preocupa-se com a maturidade global da organização, e não tanto em processos individuais.

A descrição dos níveis da análise por etapas é a seguinte (Coelho, 2009):

- **Nível 1: Inicial**. Processos são normalmente *ad hoc* e caóticos
- **Nível 2: Gerido**. Existem intenções de estabelecer a organização enquanto fornecedoras de serviços. O fornecedor dos serviços assegura que os processos são executados de acordo com a política.
- **Nível 3: Definido**. Os processos *standard* encontram-se estabelecidos consistentemente em toda a organização. Os processos são definidos em consonância com as linhas orientadoras da organização.
- **Nível 4: Quantificado e gerido**. Os fornecedores de serviço estabelecem objetivos quantitativos para qualidade e performance e usam-nos como critério para gerir os processos. Objetivos quantitativos são baseados nas necessidades dos clientes, utilizadores finais e organização.
- **Nível 5: Otimizando**. O processo é melhorado com base na análise da variação dos indicadores de qualidade e performance, com base em alterações incrementais e de inovação.

É de senso comum que a adoção do CMMI por empresas se traduz em bons resultados na redução de custos e anomalias (Coelho, 2009). Este nível de maturidade não, embora adequado e focado na entrega de serviço, não integra de forma implícita a cultura da ITIL.

2.1.4.4 Comparativo dos modelos

Os modelos de maturidade não apresentam entre eles diferenças muito significativas, todos estão dispostos por níveis, objetivos e práticas. Apesar disso, existem aspetos de cada um a ter em consideração.

O CMM é um modelo conhecido e é utilizado como base de outros modelos. É consistente, coerente e completo; contudo, é mais vocacionado para o desenvolvimento de *software* e apenas assenta na representação por etapas.

O PMF é um modelo pequeno, resumido a seis páginas no livro da ITIL, com poucos objetivos por nível. Desta forma, torna-se insuficiente esta aferição para garantir que uma organização está de facto num determinado nível de ITIL.

O CMMI *Services* é conhecido mundialmente, sendo completo, focado no serviço e contem conceitos que são facilmente transportáveis para outros standards conhecidos, como a ITIL, Cobit e outros. Descreve ambos os modelos, por fases ou por processo, sendo a abordagem sempre focada aos serviços.

Neste estudo dar-se-á maior importância ao CMMI, por ser o mais suportado e estreitamente ligado à ITIL.

2.2 Gestão da Mudança

O ser humano por natureza tende a encontrar o seu ponto de equilíbrio, dentro da sua esfera de conforto. Toda e qualquer mudança que tenha impacto na sua envolvente tende o obrigar a alterações no raio da esfera de conforto, obrigando a adaptações e reorganização do modo de ser e estar do individuo, situações essas que originam um desconforto transitório até ser encontrado o novo ponto de equilíbrio, ou reajustada a esfera de conforto.

Num ambiente organizacional este paradigma surge quando são efetuadas alterações processuais ou comportamentais obrigando, pelas boas práticas, a que as exista algum cuidado na forma como é gerida a mudança, atenuando assim os efeitos negativos da mesma. Após a importância e relevância da mudança proposta estar assegurada e comprovada, seja para benefício global da entidade ou simplesmente para melhoria e otimização de um processo ou mesmo para a obtenção de novos outputs, é imprescindível delinear o processo de mudança sustentando-o numa base de liderança e confiança.

Se a frequência da ocorrência de mudanças numa organização for elevada, a tendência em não cumprir as tarefas e planeamento estipulado para a mesma poderá tornar-se tornar um hábito, afetando consequentemente o controlo que se obterá no seu impacto na organização.

Nem sempre a mudança é positiva. Daí a importância de acompanhar e acautelar todos os passos da mudança, e reconhecer e esperar como primeira etapa a resistência à mudança. Só acompanhando e planeando a mudança é possível obter os efeitos desejados da mesma, caso contrário optando por exemplo por abordagens coercivas, os resultados finais poderão ficar aquém do esperado, e o número de problemas ou dificuldades a ultrapassar nesta forma de implementação aumentam significativamente.

2.2.1 Gestão de Mudança na ITIL

A mudança organizacional é um dos efeitos da implementação de uma *framework* para a gestão de serviços TI. Este efeito manter-se-á, não só na fase de implementação mas também nas fases subsequentes, já que o ITIL pressupõe uma constante melhoria e otimização. A ITIL defende que "Documentar" é uma das atividades mais importantes, portanto documentar os processos, as alterações e os procedimentos que afetem a organização no seu todo e em particular o serviço TI terá de ser uma tarefa a assegurar pela organização.

É importante aferir o desempenho das mudanças organizacionais, permitindo demonstrar como as novas abordagens, comportamentos e atitudes influenciam a organização, garantindo que se evite também a regressão a antigas práticas. Assim, deverão ser tidos também em consideração os comentários, as necessidades, as expectativas e as ideias recebidas ao longo de todo o processo (Ferreira, 2011).

"*The main goal of Change Management is for all the changes that need to be made to IT infrastructure and services to be performed and implemented correctly by ensuring standard procedures are followed*." (Addy, 2007)

A ITIL sustenta que a Gestão de Mudança tem de funcionar com vista a garantir que as mudanças são:

- Justificadas;
- Executadas sem pôr em perigo a qualidade do serviço de TI;
- Registadas, documentadas e classificadas;
- Testadas num ambiente de teste;

- Registadas na *Configuration Manager Database* (CMDB);
- Podem ser desfeitas através de planos de recuperação se as funções de sistema funcionarem incorretamente após implementação.

As vantagens e desvantagens deste processo estão descritas na tabela 4.

Tabela 4 - Vantagens e desvantagens da gestão de mudança na ITIL

Vantagens	**Desvantagens/Riscos**
O número de potenciais incidentes e problemas associados a cada mudança são reduzidos	Os vários departamentos têm de aceitar a autoridade da Gestão da Mudança sobre os assuntos relativos à mudança
Se a mudança tiver um impacto negativo na estrutura de TI, o processo de retornar a um estado estável e seguro é relativamente simples e rápido	As pessoas responsáveis pela Gestão da Mudança não têm um conhecimento profundo da organização tornando impossível que realizem as suas tarefas de uma forma adequada
O número de retrocessos necessários é reduzido	A unidade orgânica responsável pela gestão da mudança pode não ter as ferramentas adequadas para monitorizar e documentar todo o processo
Os custos associados a uma mudança são avaliados tornando assim possível calcular o retorno do investimento	Não existe o empenho necessário aos gestores de topo para implementar os processos de forma rigorosa
CMDB é atualizada	Demasiados processos restritivos são adotados não permitindo a ocorrência natural de inovações
As mudanças são mais facilmente aceites e a tendência de resistência é atenuada	Demasiados processos restritivos são adotados tornando o processo de gestão de mudança trivial, afetando a qualidade dos serviços

Adaptado de (Tan C. C., 2006)

2.3 Análise de Conteúdos Semânticos

Um documento é composto por um conjunto ordenado de termos classificáveis enquanto substantivos, verbos, adjetivos, pronomes, entre outros, com diferentes significados dado o contexto em que estão inseridos, podendo apresentar-se em diferentes formas e consequentemente com diferentes significados.

Para que a extração de conteúdos semânticos a partir de texto não estruturado possa ocorrer, existem um conjunto de metodologias que se baseiam na análise de raiz etimológica dos termos, frequência, proximidade no texto e enquadramento em regras gramaticais previamente definidas.

O domínio da análise de conteúdos semânticos cruza-se com a semiótica, que se de define como "*Uma teoria filosófica geral dos sinais e símbolos estabelece a relação entre uma língua artificialmente construída e outra natural. Compreende uma abordagem sintática, a semântica e a pragmática.*" (Merriam-Webster Dictionary - Semiotics)

Assim, a análise de documentos e respetivos textos compreende a interpretação da linguagem natural e sua representação num modelo pragmático de símbolos e suas relações com os respetivos significados, dando-se a estas relações o nome de semântica. À relação entre os símbolos é dado o nome de sintaxe.

Desta forma, a análise de conteúdos semânticos, à luz da perspetiva semiótica, compreende duas vertentes:

- Análise sintática do texto;
- Análise semântica do texto.

Com base nestas duas vertentes de análise, podem posteriormente ser criados sistemas que relacionam os termos, suas construções gramaticais e respetivas sintaxes, para extraírem conceitos que por sua vez se podem novamente interrelacionar. Assim, torna-se percetível que a análise terá forte influência consoante o contexto dos documentos e sua área, podendo ser por exemplo distinto extrair conhecimento de um texto relacionado com medicina comparativamente com extrair conhecimento de um texto relacionado com gastronomia. O grande desafio é portanto encontrar metodologias para extração de conhecimento, em particular nas áreas de investigação a que os documentos em estudo se relacionam.

Na tabela 5 são apresentadas as fases macro na análise e extração de conteúdos semânticos, bem como as principais metodologias aplicáveis em cada uma das fases.

Tabela 5 - Técnicas de mineração e considerações chave

Técnicas de mineração de dados	**Considerações chave**
Organização e estruturação de conteúdos	• Categorization • Classification • Clustering • Taxonomy
Processamento de texto	• Natural Language Processing (NLP) • Parsing • Parts-of-Speech (POS) Tagging • Stemming • Term Reduction • Tokenization
Análise Estatística	• Distribution • Document Indexing • Document Term Matrix (DTM) • Keyword Frequency • Term Frequency • Term Frequency - Inverse Document Frequency (TF-IDF)
Machine Learning	• Clustering • Classification • Association Rules • Predictive Modeling
Métodos de Classificação	• Naive Bayes • Support Vector Machines • k-nearest Neighbor
Avaliação do Modelo	• Precision • Recall • Accuracy • Relevance

Fonte própria

Estas técnicas e respetivas considerações-chave serão aprofundadas nos pontos seguintes tendo presente a sua aplicabilidade ao estudo em causa.

2.3.1 Gestão do conhecimento

Nas últimas décadas do século XX assistiu-se a uma profunda mudança mundial nos paradigmas económicos. O princípio da terra, capital e trabalho cedeu a posição ao novo pilar da economia – o conhecimento (Davenport & Prusak, 1998), (Drucker P. , 2006), (Nonaka & Takeuchi, 1995). Neste contexto torna-se importante para as organizações, consciencializadas do potencial que o deter do conhecimento oferece, explorar e entender como o conhecimento é realmente criado, desenvolvido e partilhado (Davenport & Prusak, 1998).

Desde a pré-história que o conhecimento começou a ser desenvolvido pelo Homem, confirmado a partir de achados arqueológicos, tendo sido mais evidente esse registo a partir da invenção da escrita. Essa partilha foi potenciada através de inovações tecnológicas dos processos de comunicação, destacando-se por exemplo a imprensa por Gutenberg, permitindo a comunicação de "um para muitos".

2.3.1.1 Dado

De acordo com Valdemar Setzer (Setzer, 1999), podemos definir "*dado como uma sequência de símbolos quantificados ou quantificáveis.*". Nesta abordagem, podemos considerar um texto como um dado, tendo como alfabeto os símbolos quantificados que constituem a sua base numerável. Desta forma, torna-se possível quantificar totalmente um texto, e armazená-lo de forma em que possa ocorrer processamento sobre o mesmo, ainda que esta representação se torne ininteligível para um leitor. Setzer, refere ainda que "*um dado é necessariamente uma entidade matemática e, desta forma, puramente sintática*", querendo destacar que os dados podem ser totalmente descritos através de representações formais e estruturais. Assim, o "dado" é objetivo e carece de significado para que possa ser transformado em "informação", e consequentemente tornar-se mais útil para o ser humano.

2.3.1.2 Informação

Informação é uma abstração informal que tende a representar algo com significado para um individuo. Esta representação pode ser através de textos, imagens, sons ou animação, e não pode ser formalizada, segundo Valdemar Setzer (Setzer, 1999), através de teoria lógica ou matemática, não sendo por isso possível processar informação

diretamente num computador, sem primeiro a reduzir previamente a dados. Uma distinção fundamental entre dado e informação é que o primeiro é puramente sintático e o segundo contém necessariamente semântica (implícita na palavra "significado" usada em sua caracterização). Desta forma, é impossível, segundo Setzer, introduzir semântica num computador, porque a máquina é puramente sintática.

2.3.1.3 Conhecimento

O conhecimento, de acordo com Davenport (Davenport & Prusak, 1998) é "*uma mistura fluida de experiência condensada, valores, informação contextual e insigh experimentado, a qual proporciona uma estrutura para a avaliação e incorporação de novas experiências e informações. Ele tem origem e é aplicado na mente dos conhecedores. Nas organizações, ele costuma estar embutido não só em documentos ou repositórios, mas também em rotinas, processos, práticas e normas organizacionais.*". Assim, os dados que representam uma informação podem ser armazenados num computador, mas a informação em si não pode ser processada quanto a seu significado, pois depende de quem a recebe, e da sua experiência adquirida anteriormente.

2.3.2 Processo de descoberta do Conhecimento

A obtenção de conhecimento, através de meios computacionais, encontra-se dividida em duas seções, nomeadamente ao processo de descoberta de conhecimento em base de dados e 2.3.2.1.3 processo de descoberta de conhecimento em texto, que se detalham em seguida.

2.3.2.1 Processo de Descoberta de conhecimento em base de dados

A extração de conhecimento existente em registos da base de dados (BD) é um problema comum em praticamente toda a área de ciência. Segundo Wives (Wives & Loh, 2006), a descoberta de conhecimento na área de ciência de computação surgiu na inteligência artificial, a qual, entre outros objetivos, se preocupava com a aquisição e armazenamento de conhecimento. Com a evolução dos sistemas de gestão de bases de dados (Korth & Silbertchatz, 1993), os pesquisadores de sistemas ou tecnologias de informação passaram a investigar novas formas de tratar informações armazenadas em bases de dados, o que originou a aparecimento de uma grande quantidade de métodos de

mineração para resolver este tipo de problemas (Alexe, Alex, Hammer, & Kogan, 2002). Estes novos métodos resultaram na criação da ferramentas que permitem moldar os dados ou agrupá-los de forma conveniente, e obter assim relações ou inferir nova informação sobre os mesmos. Alguns exemplos são as ferramentas de *Online Analytical Processing* (OLAP) e os conceitos de *Data Warehouse (DW)* (Wives & Loh, 2006).

2.3.2.1.1 Reconhecimento de padrões

De acordo com Simon Haykin (Haykin, 1999), os seres humanos são bons reconhecedores de padrões. Tal processo ocorre, na maioria das vezes, de forma impercetível e natural, dando-se exemplos de alguns cenários:

- Reconhecer um rosto familiar após envelhecimento;
- Identificar uma pessoa pela voz através de um canal de comunicação com ruido;
- Distinguir o estado de um alimento pelo seu cheiro.

O reconhecimento de padrões é formalmente definido como o processo através do qual um padrão ou sinal recebido é atribuído a uma classe de entre um número predeterminado de classes (categorias) (Haykin, 1999). Extrapolando essa analogia para os exemplos anteriores, onde o rosto, a voz e o cheiro são atribuídos a classes (categorias) específicas, torna-se assim possível reconhecer os padrões respetivos, as classes: rosto familiar, uma pessoa e o estado (qualidade) do alimento.

Apesar da para o ser humano o reconhecimento de padrões aparentar ser um processo natural, a reprodução do mesmo em seio computacional requer processos bastante complexos. Uma das formas de realizar tais tratamentos computacionalmente é com recursos à utilização de técnicas de Redes Neurais Artificiais (*Artificial Neural Network* (ARN)) (Alexe, Alex, Hammer, & Kogan, 2002).

Uma ARN reconhece padrões recorrendo a uma etapa inicial de treinamento, onde um conjunto de padrões de entrada é apresentado repetidamente à classe (categoria) à qual cada padrão pertence. Numa segunda etapa, apresentam-se novos padrões não vistos anteriormente, mas que pertencem às categorias já conhecidas, esperando-se que a ARN os classifique com base em métodos estatísticos. O reconhecimento de padrões somente

representa conhecimento caso seja facilmente compreendido pelo ser humano, útil e novo.

O reconhecimento de padrões é utilizado em processos de mineração de dados, conforme se pode ver a seguir.

2.3.2.1.2 Tipos de padrões descobertos

De acordo com J. Han (Han & Kamber, 2001), as tarefas (funcionalidades) da mineração de dados (*Data Minning* (DN)) podem ser descritivas ou preditivas. As tarefas de mineração descritivas assentam essencialmente na caracterização das propriedades gerais de uma Base de Dados (BD). As tarefas de mineração preditivas executam inferência sobre os dados existentes por forma a criar predições, ou seja, geração de novos dados. Assim, as funcionalidades de DM e tipos de padrões podem ser:

- **Descrição Classe/Conceito**: estas descrições podem ser obtidas através da caracterização dos dados ou sumarização dos mesmos, da descriminação de classes alvo com base num conjunto de classes comparativas ou de descriminação direta dos dados.
- **Análise de Associação**: trata da descoberta de regras de associação mostrando condições de atributo-valor que frequentemente são observadas num conjunto ou grupo de dados
- **Predição e Classificação**: processo que permite encontrar um conjunto de modelos para descrever ou distinguir classes ou conceitos, com o propósito de habilitar o uso de modelos de previsão de classes cujo rótulo da classe é desconhecido;
- **Análise de Agrupamentos**: é um método capaz de analisar uma série de objetos e identificar entre eles correlações e similaridades entre eles (Wives, 2004).

2.3.2.1.3 Processo de Descoberta de conhecimento em texto

Para Wives (Wives & Loh, 2006), com o advento e popularização da Internet e seus serviços, iniciou-se uma geração de um grande volume de informações não estruturadas e semiestruturadas.

A necessidade de descobrir conhecimento em tão vasto volume de informação levou ao surgimento de uma nova área de descoberta de conhecimento intitulada: KDT (*Knowledge Discovery from Texts* ou Descoberta de Conhecimento em Textos) (Wives, 2004).

Ainda segundo Wives e outros (Wives & Loh, 2006), destacam-se as seguintes formas de descoberta de conhecimento em texto:

- **Descoberta tradicional após extração**: nesta abordagem, os dados são extraídos dos textos e formatados em bases de dados estruturadas com o auxílio de técnicas de extração de informação;
- **Descoberta por análise linguística**: nesta abordagem as regras e informações podem ser descobertas através de análises linguísticas em nível léxico, morfológico, sintático e semântico;
- **Descoberta por análise de conteúdo**: nesta abordagem investiga-se os textos e apresenta-se ao utilizador informações sobre o seu conteúdo;
- **Descoberta por sumarização**: nesta abordagem utiliza-se técnicas linguísticas e extração por passagem para criar resumos;
- **Descoberta por associação entre passagens**: este tipo de técnica busca encontrar automaticamente conhecimento e informações relacionadas no mesmo texto ou em textos diferentes;
- **Descoberta por lista de conceitos-chave**: esta abordagem baseia-se na ideia de que o significado de um texto não é determinado pela sua leitura linear, mas sim, por uma análise do conjunto de elementos léxicos mais importantes (palavras-chave);
- **Descoberta de estruturas de textos**: esta abordagem baseia-se na determinação da estrutura do texto para entender o seu significado;
- **Descoberta por *clustering* (agrupamento ou aglomerados)**: procura-se separar automaticamente elementos em classes que são identificadas durante o processo (não há classes pré-definidas);

- **Descoberta por descrição de classes de textos**: esta abordagem baseia-se no facto de se ter uma classe de documentos textuais (já agrupados) e uma categoria associada a esta classe; procura-se encontrar as principais características destas classes de forma que se torne possível identificá-las e distingui-las das demais classes;
- **Descoberta por recuperação de informações**: nesta abordagem os sistemas IR (*Information Retrieval* ou Recuperação de Informação), na sua operação tradicional, contribuem para que os utilizadores obtenham novos conhecimentos;
- **Descoberta por associação entre textos**: nesta abordagem procura-se relacionar as características presentes em vários textos diferentes;
- **Descoberta por associação entre características**: nesta abordagem procura-se tipos de informações presentes em textos aplicando-se técnicas de correlação estatística (KDD);
- **Descoberta por hipertextos**: nesta abordagem, a descoberta é exploratória e experimental;
- **Descoberta por manipulação de formalismos**: nesta abordagem, utiliza-se manipulação simbólica para inferir novos conhecimentos;
- **Descoberta por combinação de representações**: nesta abordagem os textos antes de serem combinados, passam por um processo de representação interna (pressupõe a existência de dois textos);
- **Descoberta por comparação de modelos mentais**: nesta abordagem, procura-se representar documentos textuais e o estado de conhecimento do utilizador (modelo mental) através de um formalismo padrão, para, logo de seguida os comparar.

Para Loh e outros (Loh, Amaral, Wives, & de Oliveira, 2006), a técnica de descoberta de conhecimento é importante para quem necessita de trabalhar com um grande volume de informações, permitindo a descoberta de conhecimento útil e novo, geralmente implícito, minimizando a sobrecarga de informações.

Segundo (Chen¸Hsinchun & Chiang, 2012), é necessária a criação de metodologias de interação orientadas ao vocabulário, já que textos de diferentes temas podem incluir terminologias diferentes ou específicas, como por exemplo a terminologia médica ou

informática. O autor argumenta a existência de estratégias, para solução do problema do vocabulário, com uma abordagem baseada em conceitos:

- **Identificação do Vocabulário**: o mais popular meio de comunicação é através da linguagem natural. Assim, podem-se aproveitar saídas textuais para identificar o vocabulário utilizado, bem como criar um espaço de conceitos. Para tal, pode-se recorrer a técnicas de Inteligência Artificial, especificamente, processamento em linguagem natural associado a um domínio de conhecimento específico.
- **Ligação de Similaridades do Vocabulário**: neste conceito recorre-se à indexação automática de textos baseada em conceito proveniente da técnica proposta por Salton (Salton & A., 1975), conhecida como VSM (*Vector Space Model* ou Modelo de Espaço Vetorial).

Nesta técnica, identifica-se, tipicamente, a importância de termos através de cálculo de TF (*Term Frequency* ou frequência de termo no documento), DF (*Document Frequency* ou frequência do termo no conjunto de documentos) e IDF (*Inverse Document Frequency* ou frequência inversa do termo no conjunto de documentos, ou seja, os termos menos frequentes nos documentos são os mais importantes). Esta abordagem baseia-se na análise de agrupamentos e é uma extensão de VSM para a geração do espaço de conceito. Os pesos estatísticos entre termos indicam a sua forte relevância ou associação.

2.3.3 Recuperação de informação (*IR – Information Retrieval*)

O termo IR (*Information Retrieval* ou Recuperação de Informação) foi desenvolvido por Calvin Moore entre 1948 e 1950, sendo um campo de pesquisa interdisciplinar, baseado em muitas áreas. Dada a sua abrangência o autor não é muito compreendido, sendo frequentemente abordado sob perspetivas diferentes. Moore posiciona-se na junção de muitos campos já estabelecidos, tais como: Psicologia Cognitiva, Arquitetura da Informação, Projeto da Informação, Comportamento da Informação Humana, Linguística, Semiótica, Ciência da Informação, Ciência da Computação, Biblioteconomia e Estatística (Baeza-Yates & Neto, 1999).

A área de IR refere-se aos sistemas automáticos de recuperação de informação que permitem encontrar documentos relevantes em função de uma necessidade de informação específica. Noutra definição, é a área da Ciência da Computação que se preocupa com a seleção, num universo de documentos disponíveis, do conjunto de documentos relevantes para uma necessidade específica de informação.

Baeza-Yates (Baeza-Yates & Neto, 1999) define formalmente os modelos IR focando a abordagem quantitativa, como:

Um modelo de recuperação de informação é uma quádrupla $[D, Q, F, R(qi, dj)]$ onde:

1. D é um conjunto composto de visões lógicas (ou representações) para o documento na coleção;
2. Q é um conjunto composto de visões lógicas (ou representações) para as necessidades de informações. Tais representações são conhecidas como consultas;
3. F é uma modelo que permite definir representações de documentos, perguntas, e seus relacionamentos;
4. $R(qi, dj)$ é uma função de ranking que associa um número real com uma consulta qi (de Q) e uma representação do documento di (de D). Tal ranking define uma ordenação entre os documentos no que diz respeito à consulta qi.

Os modelos de IR categorizam-se, na ótica das tarefas executadas, em três classes, nomeadamente: pesquisa, filtragem e navegação.

Desta forma, a tarefa de pesquisa é interativa mediante a necessidade de informação, que é esporádica e baseada em consultas. A filtragem é baseada em configurações onde as necessidades de informação são de caráter permanente e baseadas em perfis específicos. Finalmente, a navegação é interativa, sendo a necessidade de informação indefinida e a formulação é baseada no percurso (Baeza-Yates & Neto, 1999).

Os modelos de IR podem ser categorizados como quantitativos ou dinâmicos. Na taxonomia dos modelos quantitativos destacam-se os modelos clássicos, compostos pelos modelos booleanos, vetoriais e probabilísticos. No modelo booleano, os documentos e as consultas são representados por conjuntos de termos índices baseados na teoria de conjuntos. No modelo vetorial, documentos e consultas são representados como vetores em um espaço t-dimensional em um modelo algébrico. No modelo probabilístico, os documentos e as consultas são representados com base na teoria de probabilidade.

Com o decorrer do tempo, foram propostos modelos alternativos aos modelos clássicos, assentes em abordagens algébricas, como por exemplo vetor generalizado, indexação semântica latente e redes neurais.

2.3.4 Linguística Computacional

A linguística computacional é um campo multidisciplinar para tratamento da língua natural com base em conhecimentos estatísticos e/ou com base em regras da linguagem (padrões linguísticos) de uma perspetiva computacional.

2.3.4.1 Processamento de linguagem natural

NLP (*Natural Language Processing*) é um subcampo da Inteligência Artificial e da Linguística dedicado ao estudo dos problemas de automação do processo de geração e entendimento da linguagem humana natural. O objetivo principal do NLP é conseguir uma melhor compreensão da língua natural, através do uso de computadores. Pode-se anexar ao NLP a criação de técnicas para o processamento rápido de textos. Neste contexto, emprega-se técnicas que vão desde manipulação de elementos textuais até ao processamento automático de consultas.

Destacam-se de seguida alguns dos problemas que os pesquisadores de NLP pretendem resolver:

- **Segmentação da Fala**: na maioria das línguas faladas, os sons podem ser representados por letras sucessivas que se misturam foneticamente. Assim, a conversão de sinal analógico para caracteres discretos pode ser um processo muito complexo.

- **Segmentação de Texto**: em algumas línguas escritas, como por exemplo o Chinês ou Japonês, podem não existir delimitadores de palavras específicos. Assim, todo texto significativo que é analisado gramaticalmente requer geralmente a identificação de limites de palavra, que é frequentemente uma tarefa não trivial.
- **WSD *(Word Sense Disambiguation ou Desambiguação de Sentido de Palavra)***: Muitas palavras podem ter mais de um significado consoante o contexto onde estão inseridas. Assim torna-se necessário aplicar técnicas computacionais que permitam identificar adequadamente e utilizar o significado em cada contexto (Das Graças Volpe Nunes & Specia, 2004).
- **Ambiguidade Sintática**: a gramática da língua natural (como por exemplo, o Português) é ambígua, ou seja, existem múltiplas possibilidades de árvores de análise para cada sentença. Escolher a mais apropriada, usualmente, requer informação de contexto e semântica. Os componentes específicos do problema da ambiguidade sintática incluem a desambiguação do limite da sentença (Das Graças Volpe Nunes & Specia, 2004).

O processamento estatístico de língua natural usa métodos estocásticos, probabilísticos e estatísticos para resolver algumas dificuldades relacionadas às ambiguidades que, quando processadas através de gramáticas formais, podem atingir milhares ou milhões de análises possíveis.

2.3.4.2 Linguística de Corpus

Sardinha (Sardinha, 2004) define a linguística de corpus como "*Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade de uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.*"

A linguística de corpus realiza uma série de processos sendo, um dos mais importantes, a **etiquetagem**. Segundo Sardinha, o etiquetador serve para inserir automaticamente ou

de forma semiautomática (interativa) no corpus informações sobre as palavras, consistindo nos seguintes tipos:

- **Morfossintática (*Part of Speech* ou POS)**: também denominada morfológica, consiste na marcação das classes gramaticais (como por exemplo: substantivo, verbo, adjetivo etc.) de cada palavra.
- **Sintática (*Parsing*)**: consiste na identificação da estrutura sintática (como por exemplo: sintagma nominal, sintagma verbal, etc.) de cada frase.
- **Semântica (*Semantic* ou rotulagem de sentidos)**: consiste na definição do sentido ou categoria semântica de cada palavra (como por exemplo: $casa = moradia$, $alicate = ferramenta$ etc.). Essas categorias são relativamente genéricas, podendo ser constituídas, por exemplo, de rótulos ou etiquetas que denominam a área ou domínio de determinada palavra num contexto ou outras etiquetas ontológicas, como "humano", "animado" etc. (Das Graças Volpe Nunes & Specia, 2004).
- **Discursiva (*Discourse*):** consiste na marcação de características como referentes anafóricos, tópicos ou marcadores discursivos (Sardinha, 2004).

A etiquetagem de corpus consiste da inserção de informações referentes a cada unidade de texto (morfológica, sintática, semântica e discursiva). Uma das mais utilizadas na linguística de corpus é a morfossintática ou POS *tagging*, que se baseia na explicitação de classes gramaticais de cada palavra. Este processo auxilia na desambiguação lexical (Sardinha, 2004).

2.3.5 Preparação de Corpus Textuais

2.3.5.1 Itemização

A itemização é o processo pelo qual é possível obter todas as palavras usadas num texto, e é também referenciado como Análise Léxica. Este processo consiste na subdivisão do documento textual num conjunto de palavras removendo para isso todas as marcas de pontuação, divisão ou outros caracteres não textuais por espaços, permitindo apenas a existência de um espaço único entre dois símbolos (ou *token*) (Hotho, Nürnberger, & Paaß, 2005). O resultado do processo de itemização é um conjunto de palavras diferentes, denominado de dicionário da coleção de documentos

(Hotho, Nürnberger, & Paaß, 2005). Após o processo de itemização estar concluído, a representação do documento textual poderá ser posteriormente usada em processamentos adicionais, como por exemplo a etiquetagem. De uma forma geral, a itemização consiste na separação de unidades ortográficas (Sardinha, 2004).

Segundo Frakes (Frakes & Baeza-Yates, 1992), a análise léxica nos sistemas IR exige algumas formas de tratamento especiais, destacando-se:

Frequência Absoluta dos Termos: Seja D um conjunto de documentos, denotado por $D = \{d_1 \dots d_m\}$ e T o dicionário, isto é, o conjunto de todos os termos diferentes em D, denotado por $T = \{t_1 \dots t_m\}$, a Frequência Absoluta dos Termos $t \in T$ num documento $d \in D$ é obtida de $tf(d, t)$.

Denota-se assim o Vetor de Termos por $\overrightarrow{t_d} = \left(tf(d, t_1), \dots, tf(d, t_m)\right)$ (Hotho, Nürnberger, & Paaß, 2005)

Para a melhor escolha de termos possível, técnica referida em pontos seguintes desta dissertação, torna-se necessária a noção de centróide de um conjunto X de vetores do termo, isto é, o Centróide do Vetor de Termo.

Assim, seja X o vetor de termos, define-se o centróide como o valor médio do vetor de termos, denotado por $\overrightarrow{t_x}$, onde $\overrightarrow{t_x} = \frac{1}{|X|} \sum_{\overrightarrow{t_d} \in X} \overrightarrow{t_d}$ (Hotho, Nürnberger, & Paaß, 2005), sendo $|X|$ o número de elementos auxiliares do vetor de termo $\overrightarrow{t_d} \in X$ e X o conjunto de vetores de termos.

Comummente o dicionário resultante apresenta uma dimensão demasiado grande para representar a coleção de documentos, o que se traduz tipicamente num novo problema a resolver, sobretudo quando se pretendem utilizar processos de indexação. Para resolver este problema são necessárias técnicas e métodos adicionais para minimizar e reduzir o dicionário ao conjunto mínimo representativo, assegurando que o mesmo continue a manter a representatividade do conjunto de documentos após redução.

2.3.5.2 Filtragem

A designação filtro refere-se frequentemente a um dispositivo ou algoritmo que permite extrair, de um conjunto de dados ruidosos ou de maior dimensão, informações de um determinado interesse ou dimensão (Haykin, 1999).

No contexto desta dissertação e no âmbito da preparação de corpus textuais, filtragem é o método ou ação que remove palavras do dicionário ou até mesmo documentos em análise. Um dos métodos principais para a filtragem é a filtragem através de *stop words*. *Stop words*, termo definido por Hans Peter Luhn tem por base a remoção de palavras com pouco ou nenhum conteúdo semântico, de tal forma que a sua remoção provoque o mínimo possível de perdas no conjunto *bag-of-words* que representa o documento (Hotho, Nürnberger, & Paaß, 2005).

Este método não é isento de inconvenientes, tal como refere Joaquim Ferreira da Silva (da Silva, 2003), já que a eliminação por acarretar dois problemas. O primeiro prende-se com o facto de geralmente serem utilizados filtros linguísticos para eliminar as *stop words*, o que constitui uma limitação à aplicabilidade do método a diferentes idiomas. O segundo lugar, as *stop words* a eliminar podem ser informativas, alterando ou mesmo fazendo com que o *bag-of-words* deixe de ser representativo do conjunto de documentos.

Para atenuar estes possíveis problemas, uma abordagem complementar muito utilizada em processos de IR é a eliminação de palavras que ocorrem com muita frequência, uma vez que por esse facto têm pouca relevância para estabelecer diferenças entre os documentos (Baeza-Yates & Neto, 1999). Frequentemente a remoção de *stop words* é realizada através de listas de palavras vazias[2], *stoplist*, ou dicionário negativo. Tradicionalmente, a *stoplist* são compostas pelas palavras mais frequentes, geralmente as que apresentam uma frequência de 80% ou mais no texto (Pardo & das Graças Volpe Nunes, 2003), (Baeza-Yates & Neto, 1999).

[2] Palavras gramaticais, funcionais de classe fechada ou instrumentais: são palavras de categoria fechada (pronome, artigo, preposição, conjunção; algumas classificações incluem aqui verbos auxiliares e modais (Sardinha, 2004))

Outra abordagem metodológica de minimização do dicionário são os métodos e técnicas relacionadas com a fusão ou redução das palavras às suas formas básicas, apresentados a seguir.

2.3.5.3 Lematização/Stemming

Os métodos de **lematização** assentam na redução das formas e tempo dos verbos para o seu infinitivo respetivo, e as formas dos substantivos para o singular. Para que esta redução seja viável, as formas das palavras têm de ser bem conhecidas, e a parte do discurso ou *Part of Speech* ou a classe gramatical de cada palavra no texto tem de ser atribuída. Assim, este método pressupõe a execução antecipada de um processo de etiquetagem do texto. Este é um processo que consome geralmente muito tempo, o que pode originar a existência de erros de corte (quando se recorre por exemplo a árvores sintáticas). Desta forma, a lematização é normalmente menos utilizada em detrimento da técnica de *stemming* (Jacquemin, 1996).

Stemming (ou Redução Gramatical) é um processo que permite reduzir palavras derivadas para a sua forma base ou raiz linguística. A base não necessita ser idêntica à raiz morfológica da palavra; é geralmente suficiente que as palavras relacionadas tenham a mesma base, mesmo que esta base não seja uma raiz válida em termos linguísticos.

Segundo Wives (Wives, 2004), o *stemming* é uma técnica de padronização de vocabulário que permite identificar termos similares e reduzi-los a um único radical, representando-os num único termo ou forma, eliminando suas diferenças morfológicas ou lexicais.

Existem diversas técnicas de *stemming* com divergências ao nível do desempenho, exatidão e forma como determinados obstáculos são superados. Destacam-se os seguintes algoritmos:

- **Força Bruta**: O termo é originado de um conceito na área de pesquisa em Inteligência Artificial e Matemática para resolver um problema (Roche, 2003).
  Os *stemmers* de força bruta empregam uma tabela de procura (*lookup*) que contém as relações entre formas raiz ou *stem* e formas flexionadas. Para cada palavra, a tabela de raiz é consultada para encontrar uma flexão que combina. Se uma combinação for encontrada é retornada a forma associada à raiz;
- **Descarte de sufixo**: os algoritmos de descarte de sufixo são baseados em listas, tipicamente menores que tabelas *lookup*, que são as regras que são armazenadas e tratadas pelo algoritmo para a obtenção da redução;
- **Estocásticos**: os algoritmos estocásticos recorrem ao estudo das probabilidades para identificar a forma raiz de uma palavra. Os algoritmos estocásticos são treinados ("aprendem")[3] através de uma tabela de formas raiz (ou radical). Estas formas raiz, possuem relações de flexões de formas para desenvolver um modelo probabilístico. Este modelo é expresso, tipicamente, na forma de regras linguísticas complexas, de natureza similar aos algoritmos de descarte de sufixo e lematização.
- **Abordagem Híbrida**: as abordagens híbridas usam uma ou mais das abordagens descritas anteriormente. Um exemplo simples é um algoritmo de árvore do sufixo que consulta primeiramente uma tabela *lookup* usando a força bruta. Entretanto, em vez de tentar armazenar o conjunto inteiro das relações entre palavras de um dado idioma, a tabela do *lookup* é mantida em dimensão reduzida e usada somente para armazenar exceções frequentes. Se a palavra não estiver na lista de exceções, deve-se aplicar o descarte ou lematizar o resultado.

2.3.5.4 Seleção de Termos Índices

Para uma maior redução do número de palavras do dicionário pode recorrer-se a algoritmos para seleção de palavras-chave ou de indexação, onde apenas as palavras

[3]Recorrendo a técnicas de ARN que possibilitam a aprendizagem

selecionadas são usadas para descrever os documentos pertencentes aos *corpus* em análise.

Um método simples para a seleção de palavras-chave assenta na sua extração com base na sua entropia, tal como descrito na teoria de Claude Shannon melhorada por outros (Sannon, 1948) (Haykin, 1999). Nesta teoria, para cada palavra t no vocabulário a entropia pode ser computada utilizando-se as Equações seguintes definidas por Lochbaun e outros, citadas em (Hotho, Nürnberger, & Paaß, 2005):

$$W(t) = 1 + \frac{1}{log(|D|)} \sum_{d \in D} P(d,t)\ log_2\ P(d,t)$$

com

$$P(d,t) = \frac{tf(d,t)}{\sum_{l=1}^{n} tf(d,t)}$$

onde, $W(t)$ é a entropia de um termo t e os demais termos são conforme definição de Frequência Absoluta de Termos no ponto 2.3.5.1.

A entropia dá assim informação sobre quão representativa é uma palavra no momento da recuperação através da busca por palavras-chave ou termos índices do documento em análise. Pode observar-se que as palavras que ocorrem em muitos documentos possuem baixa entropia. A entropia deve portanto ser vista como uma medida da importância de uma palavra no contexto de um dado domínio de conhecimento (Hotho, Nürnberger, & Paaß, 2005).

2.3.5.5 Modelo de Espaço Vetorial

O VSM (*Vector Space Model* ou Modelo de Espaço Vetorial), defendido por Salton (Salton & A., 1975), foi originalmente introduzido para indexação em sistemas IR, assentando numa abordagem puramente estatística. Apesar de possuir uma estrutura de dados simples, que não usa nenhuma informação semântica explícita, – o que se traduz numa economia no custo computacional - o VSM permite uma análise muito eficiente

de grandes coleções de documentos. Atualmente é utilizado em abordagens de KDT, bem como, em muitos sistemas IR (Hotho, Nürnberger, & Paaß, 2005).

O VSM representa os documentos com recurso a vetores definidos no espaço m-dimensional, ou seja, cada documento d é descrito por um vetor numérico de características $\overrightarrow{w(d)} = \big(x(d,t_1), \dots, x(d,t_m)\big)$.

No vetor de características $\overrightarrow{w(d)}$, cada elemento $x(d,t_i)$ refere-se à ocorrência do termo t_i no documento d. Desta forma, os documentos podem ser comparados pelo uso de operações vetoriais simples ou através de codificação de consultas de termos baseadas na similaridade vetorial. O vetor de consulta pode então ser comparado com cada documento obtendo-se uma lista de resultados através da qual é possível computar a sua similaridade (Hotho, Nürnberger, & Paaß, 2005).

Importa assim encontrar uma codificação apropriada para o vetor de características que representa o documento, tal como refere Hotho (Hotho, Nürnberger, & Paaß, 2005). Cada elemento do vetor representa geralmente uma palavra (ou um grupo de palavras) do documento da coleção, ou seja, o tamanho do vetor é definido pelo número de palavras (ou grupos) da coleção completa do *corpus*. A forma mais simples de codificar a representação de um documento, passa pela utilização de vetores binários, ou seja, com valor 1 (um) caso o termo esteja presente no documento, ou 0 (zero) caso não esteja. Tal codificação resultará em comparações booleanas simples, o que melhora o desempenho computacional.

Nesta base booleana, e para melhoria da performance na análise, pode recorrer-se ao uso de pesos de forma a refletir a importância de uma dada palavra num documento, atribuindo pesos maiores aos termos que são usados com maior frequência e em documentos relevantes, mas raramente na coleção inteira de documentos. Desta forma, um peso $w(d,t)$ para um termo t no documento d é computado pela frequência do termo, denotado por $tf(d,t)$ multiplicada pela frequência inversa do termo, denotada por $idf(t)$, que descreve especificamente o termo dentro do *corpus* (Hotho, Nürnberger, & Paaß, 2005).

$$idf(t) = \log({}^{N}/_{\eta_t})$$

Além da frequência relativa do termo e a frequência inversa do documento - definida como $idf(t) = \log({}^{N}/_{\eta_t}) -$ para assegurar que todos os documentos possam ter representatividade semelhante independentemente do seu comprimento ou tamanho, importa introduzir um fator e normalização do comprimento (Hotho, Nürnberger, & Paaß, 2005):

$$s\omega(d,t) = \frac{tf(d,t)(\log {}^{N}/_{\eta_t})}{\sqrt{\sum_{j=1}^{m} tf(d,t_j)^2 (log {}^{N}/_{\eta_t})^2}}$$

onde, N é o tamanho da coleção de documentos D, ou seja, o corpus em análise, e ηt é o número de documentos em D que contêm o termo.

De acordo com o esquema de peso, um documento d é definido por um vetor de pesos de termos $\overrightarrow{w(d)} = \big(x(d,t_1), \dots, x(d,t_m)\big)$ e a similaridade S de dois documentos d_1 e d_2 (ou a similaridade de um documento e um vetor de consulta) pode ser calculada através do produto escalar dos vetores:

$$S(d_1,d_2) = \sum_{k=1}^{m} \omega(d_1,t_k).\omega(d_2,t_k)$$

onde, $S(d_1,d_2)$ representa a similaridade entre os vetores e $\omega(d,t)$ os vetores de pesos. Uma medida frequentemente utilizada é a da Distância Euclidiana [109].

Pode-se calcular a distância entre dois documentos $d_1, d_2 \in D$ da forma apresentada na equação a seguir:

$$dist(d_1,d_2) = \sqrt{\sum_{k=1}^{m} |\omega(d_1,t_k) - \omega(d_2,t_k)|^2}$$

onde, $dist(d_1,d_2)$ é a distância entre os documentos e $\omega(d,t)$ os vetores de pesos.

Entretanto, a distância euclidiana somente pode ser usada para vetores normalizados, pois, documentos de tamanhos diferentes podem resultar em pequenas distâncias entre os documentos que compartilham poucas palavras e os que possuem mais palavras em comum e poderiam ser considerados como mais similares.

Para Hotho (Hotho, Nürnberger, & Paaß, 2005), quando os vetores estão normalizados, o produto escalar não é muito diferente da medida euclidiana, desde que os vetores $\vec{x}$ e $\vec{y}$ estejam conforme se apresenta na equação seguir:

$$\cos\varphi = \frac{\vec{x}\vec{y}}{|\vec{x}|.|\vec{y}|} = 1 - \frac{1}{2}d^2\left(\frac{\vec{x}}{|\vec{y}|}, \frac{\vec{y}}{|\vec{y}|}\right)$$

Serão apresentados de seguida alguns métodos de mineração do texto.

2.3.6 Métodos de mineração de texto

Após a execução da etapa de pré-processamento, com recurso às técnicas de etiquetador morfossintático (POS), segmentação de texto, WSD (*Word Sense Disambiguation*) e análise gramatical de sentenças (*parsing*), o processo de mineração de texto tem a seu dispor um conjunto de técnicas com finalidades específicas que possibilita a extração de conhecimentos nos *corpus* em análise.

Este ponto visa abordar algumas técnicas relacionadas com a KDT (*Knowledge Discovery in Text*), tais como a classificação, o agrupamento e a extração de informações.

A mineração de grandes coleções de documentos inicia-se normalmente com a realização de uma preparação através de um processo de pré-processamento, onde se armazenam informações numa estrutura de dados, acautelando que essa estrutura seja apropriada a posteriores processamentos.

O termo "documento" foi apresentado por um estudo de Michael Bukland, citado em (Wives, 2004), que enumerou as seguintes definições do mesmo:

- "*Qualquer expressão do pensamento humano*."

- "*Qualquer base material capaz de estender nosso conhecimento, que seja disponível para estudos ou comparação, pode ser documento.*"
- "*Qualquer fonte de informação, em formato material, capaz de ser utilizada para referência ou estudo ou como uma autoridade.*"
- "*Um documento é uma evidência que suporta um fato. [...] qualquer signo físico ou simbólico, preservado ou registro, com o intuito de representar, reconstruir ou demonstrar um fenômeno físico ou conceitual é um documento.*"

No contexto desta dissertação, documento será todo objeto eletrónico que possuir características textuais, tais como: arquivos TXT, DOC, PDF, CSV ou similares quando estes possuírem textos, bem como registos extraídos de uma base de dados.

Existem diversas abordagens que explorar as estruturas lexicais, sintáticas, semânticas e pragmáticas do texto, bem como abordagens estatísticas. Na sua maioria, estas abordagens são baseadas na ideia de o documento textual poder ser representado por um conjunto de palavras, o *bag-of-words*. Dentro destas palavras procura-se definir a importância de cada palavra dentro de cada documento, sendo geralmente esta representação feita com recurso a um vetor, onde a cada palavra é atribuído e armazenado um valor numérico referente à sua importância.

As abordagens atualmente predominantes são baseadas na ideia da montagem de conjuntos *bag-of-words*, podendo-se destacar a VSM (*Vector Space Model* ou Modelo do Espaço do Vetorial), o Modelo Probabilístico e o Modelo Lógico (Hotho, Nürnberger, & Paaß, 2005).

2.3.6.1 Classificação

A classificação do texto, segundo T. Michell citado em (Hotho, Nürnberger, & Paaß, 2005), visa atribuir classes predefinidas aos documentos textuais. Para Wives (Wives, 2004), na classificação, as características dos objetos já são conhecidas e é possível estabelecer uma descrição para eles, ou seja, o processo de classificação consiste em analisar um objeto e associá-lo a uma das classes previamente definidas.

Formalmente, e segundo Hotho (Hotho, Nürnberger, & Paaß, 2005), a tarefa de classificação em mineração de dados inicia-se com um conjunto de treino $D = (d_1, \dots, d_n)$ de documentos já etiquetados (marcados) e uma classe $L \in \mathbb{L}$, como por exemplo, redes informáticas, religião, gastronomia, etc. A tarefa é então determinar um modelo de classificação que pode atribuir corretamente uma classe a um novo documento d a um domínio de conhecimento específico em $\mathbb{L}$, conforme se pode ver na equação seguinte:

$$f: D \rightarrow \mathbb{L} \qquad f(d) = L$$

Bastos (Bastos, 2006) defende duas condições básicas para a classificação de textos:

- As classes de classificação devem existir previamente, isto é, devem ser conhecidos o número de classes e sua identificação (significado).
- Deve existir conhecimento sobre as classes ou sobre elementos que permitam decidir onde alocar novos elementos.

Os processos de classificação podem também, à semelhança dos sistema de IR, ser alvo de medidas de precisão e abrangência. A precisão avalia a eficácia do modelo, isto é, a capacidade com que o modelo toma decisões (classificações) acertadas. A abrangência afere o grau com que o modelo recupera ou não documentos relevantes na coleção. Podemos, mais simplesmente dizer que, a precisão avalia o quanto o modelo acerta. A abrangência avalia o quanto o modelo contabiliza (Rezende, 2003).

A precisão e abrangência podem ser calculadas pelas equações seguintes (Rezende, 2003):

$$\textbf{\textit{Precisão}} = \frac{\textit{número de itens relevantes recuperados}}{\textit{número total de itens recuperados}}$$

$$\textbf{\textit{Abrangência}} = \frac{\textit{número de itens relevantes recuperados}}{\textit{número de itens relevantes na coleção}}$$

Estas duas medidas estão, com algum grau de associação, relacionadas na maioria dos classificadores. Se os documentos tiverem um grau de associação à classe alvo elevado,

a precisão é alta. Entretanto, se muitos documentos relevantes forem negligenciados a abrangência é baixa.

A precisão e a abrangência são, na prática, medidas inversamente proporcionais, pois, quando a precisão aumenta a abrangência diminui e vice-versa. Neste caso, utiliza-se a $medida - F$ para medir o desempenho de ambos e determinar a qualidade do processo de classificação. A equação seguinte calcula esta medida (Hotho, Nürnberger, & Paaß, 2005):

$$medida - F = \frac{2}{\frac{1}{abrangência} + \frac{1}{precisão}}$$

Dar-se-ão de seguida exemplos de alguns métodos associados à tarefa de classificação de documentos textuais.

2.3.6.1.1 Seleção de Termos Índices

Após as tarefas de processamento de documentos textuais, é comum a obter-se um número elevado de palavras diferentes. Assim, e para organizar e aceder a estes documentos, torna-se imprescindível a seleção dos termos mais informativos para a composição do conjunto de termos índices. Desta forma, a tarefa de classificação, dada a sua complexidade, obriga a uma redução do número destes termos para permitir o seu manuseio. Para este feito é usualmente aplicada uma técnica de medida de importância aos termos, destacando-se por exemplo a medida de ranking denominada Ganho de Informação (Hotho, Nürnberger, & Paaß, 2005) (Toussi, 2014), onde, para um termo t_j, o $IG(t_j)$ (ou Ganho de Informação) pode ser definido pela equação seguinte (Hotho, Nürnberger, & Paaß, 2005):

$$IG(t_j) = \sum_{c=1}^{2} p(L_c) log_2 \frac{1}{p(L_c)} - \sum_{m=1}^{1} p(t_j = m) \sum_{c=1}^{2} p(L_c | t_j = m) log_2 \frac{1}{p(L_c | t_j = m)}$$

Na fórmula, $p(L_c)$ caracteriza uma parte dos documentos treinados[4] com classes L_1 e L_2, $p(t_j = 1) e\ p(t_j = 0)$ representa o número de documentos que possuem ou não o termo t_j e $p(LC|t_j = 1)$ é a probabilidade condicional das classes L_1 e L_2, se o termo t_j está contido no documento ou se espera que esteja contido nestes (Hotho, Nürnberger, & Paaß, 2005). Assim, a fórmula mede a utilidade de t_j para predizer a classe L_1 do ponto de vista da Teoria da Informação de Claude Shannon (Sannon, 1948). Pode-se determinar o $IG(t_j)$ para todos os termos e remover aqueles com menor ganho de informação.

2.3.6.1.2 Classificação de Naive Bayes

O classificador *Naive Bayes*, baseado no teorema de *Bayes*, tem por ideia base usar a junção das probabilidades das palavras e categorias para estimar as probabilidades das categorias de um novo documento (Bastos, 2006).

O algoritmo calcula, usando a regra de *Bayes*, a probabilidade à posteriori de um documento pertencer a classes diferentes, associando-o depois à classe cuja probabilidade é a mais elevada (Bastos, 2006).

A parte ingênua (*naive*) do algoritmo classificador *Naive Bayes* assenta na suposição de independência das características da palavra, ou seja, é assumido que o efeito das características de uma palavra, cuja probabilidade condicional está associada a uma categoria, é independente das características de outras palavras daquela categoria (Bastos, 2006). Assim, a probabilidade conjunta pode ser calculada como o produtório das probabilidades individuais conforme se pode observar nas equações abaixo.

Este classificador inicia o processo com a suposição de que um termo t, de um documento d_i tenha sido gerado por um mecanismo probabilístico. Supõe que a classe $L(d_i)$, de um documento d_i, tenha alguma relação com as palavras que surgem no documento. Pode assim ser descrito pela distribuição condicional $p(t_1, \dots, t_{n_i}|L(L_i)$ de n_i palavras de classes obtidas pelas equações *Bayesianas* seguintes (Hotho, Nürnberger, & Paaß, 2005):

[4] Refere-se à técnica de treinamento, que fixa a classificação de um conjunto de documentos

$$p(L_c|t_1, \dots, t_{n_i}) = \frac{p(t_1, \dots, t_{n_i}|L_c)p(L_c)}{\sum_{L\in\mathbb{L}} p(t_1, \dots, t_{n_i}|L_c)p(L)}$$

$$p(t_1, \dots, t_{n_i}|L_c) = \prod_{j=1}^{n_i} p(t_j|L_c)$$

Onde, $p(t_j|L_c)$ é a estimativa das palavras em cada classe dada pela frequência relativa no documento pertencente ao conjunto de treino, as quais, são rotuladas com L_c.

2.3.6.1.3 Classificação Nearest Neighbor

O algoritmo de classificação *K-Nearest Neighbor* ou K-NN, é muito utilizado na categorização de textos através da aprendizagem baseada na similaridade (Bastos, 2006).

Neste algoritmo, a classe de um novo documento é determinada pelo cálculo da similaridade entre o documento teste e os exemplos individuais ou agregados do conjunto de treino, e a determinação da distribuição da classe dos exemplos mais próximos ou agregados (Rezende, 2003).

Na fase de classificação são calculadas para a amostra teste as similaridades e respetivas distâncias do novo vetor a todos os vetores armazenados, sendo posteriormente selecionadas as K amostras mais próximas e descartadas as restantes. A melhor escolha para o valor de K varia em função dos dados em análise. De uma forma geral, valores maiores para K reduzem o ruído, porém, criam limites entre classes menos distintas (Bastos, 2006).

2.3.6.1.4 Árvore de Decisão

Uma descrição de árvore de decisão, segundo Crawford (Bastos, 2006), é um modelo preditivo representado por um gráfico em forma de árvore contendo as decisões a serem tomadas e a suas possíveis consequências (ex.: custo, risco, etc.) e pode ser usada para criar um plano com vista a alcançar um objetivo. Permite desta forma gerar um mapeamento de observações sobre um determinado item e obter conclusões sobre o seu

valor-alvo. A abordagem da sua construção é *top-down*, assente num algoritmo do tipo dividir para conquistar e pode ser de dois tipos:

- **Árvores de Regressão**: são aplicadas a variáveis contínuas. Quando se utiliza árvore de regressão para se predizer o valor da variável-alvo, o valor principal dessa variável (que vai para um nó folha da árvore) é o valor estimado.
- **Árvores de Classificação (ou *Classification Tree*)**: são aplicadas a variáveis discretas. Cada nó folha contém um rótulo que indica a classe predita para um determinado conjunto de dados. Neste tipo de árvore podem existir dois ou mais nós folha com a mesma classe.

2.3.6.1.5 Suport Vector Machines

O classificador SVM (*Support Vector Machines* ou Máquina de Suporte Vetorial) é obtido pela junção de dois tipos de classificadores de documentos: hierárquico e não hierárquico. O SVM usa uma estrutura de dados baseada numa em árvore de categorias (discreta), com poucos níveis, idealmente dois, e os resultados das buscas estão presentes nas folhas desta árvore (Bastos, 2006).

Neste classificador, um documento d é representado por um vetor $(t_{d_1}, \dots, t_{d_n})$ de contabilização dos pesos das palavras. Um único SVM pode somente separar duas classes, uma positiva L_1 (indicada por $y = +1$) e uma negativa L_2 (indicada por $y = -1$). No espaço do vetor de entrada, um hiperplano pode ser definido com $y = 0$ conforme a equação linear seguinte (Hotho, Nürnberger, & Paaß, 2005).

$$y = f(\overrightarrow{t_d}) = b_0 + \sum_{c=1}^{n} b_i t_{d_c}$$

Os documentos são divididos em dois conjuntos, um definido como base de treino e outro como teste. Os parâmetros b_i são ajustados em função da distância ao hiperplano. Um novo documento com o vetor de termos $\overrightarrow{t_d}$ é classificado em L_1 se o valor de $f(\overrightarrow{t_d}) > 0$ e em L_2 no caso contrário. A base de treino é usada para o algoritmo de classificação obter as características das categorias da coleção. A base de teste valida o desempenho do classificador, determinando as categorias às quais os novos documentos

pertencem. Na fase de análise, o desempenho do algoritmo é medido de acordo com o resultado obtido na classificação original dos documentos, usualmente via processo de classificação manual (Bastos, 2006).

2.3.6.2 Agrupamentos

Para Jiawei Han e outros (Han & Kamber, 2001), Análise de Agrupamentos (ou *clustering*) consiste no processo de agrupar um conjunto de objetos (físicos ou abstratos) em classes similares de objetos referenciadas como agrupamento.

Para Bastos (Bastos, 2006) a técnica de *clustering* consiste em:

- Atribuir grupos homogéneos:
- Maximizar a similaridade de objetos dentro de um mesmo *cluster*;
- Maximizar a não similaridade de objetos entre *clusters* distintos.
- Atribuir uma descrição para cada *cluster* formado.

A base deste processo assenta na Hipótese de Agrupamento de Rijsbergen, citado em (Wives, 2004), e que refere que objetos semelhantes e relevantes para um mesmo assunto tendem a permanecer num mesmo *cluster*.

Para Jiawei Han e outros (Han & Kamber, 2001), existem os seguintes métodos de *clustering*:

- **Método do Particionamento**: cria-se um número, aleatório, de K partições ajusta-se os agrupamentos através de processo interativo.
- **Método Hierárquico**: cria-se uma decomposição hierárquica, através de processos *top-down* (divisivo) ou *bottom-up* (aglomerativo).
- **Método Baseado em Densidade**: cria-se agrupamentos com base na noção de densidade.
- **Método Baseado em Grade**: quantifica-se os objetos em espaço finito de células que formam grades.

Relativamente aos algoritmos de agrupamentos, destacam-se os seguintes: Algoritmo *K-Means*, algoritmo *K-Medoids*, algoritmo *Bisecting-K-Means*, algoritmos SOM (*Self-*

Organizing Maps ou Mapas Auto-Organizáveis) e algoritmos de EM (*Expectation–maximization* ou maximização de expetativa) (Han & Kamber, 2001).

2.3.6.3 Extração de informações

A análise de linguagem natural contém usualmente muitas informações que não podem ser facilmente analisadas por processos computacionais.

Assim, a IE (*Information Extraction* ou Extração de Informação) é uma área de pesquisa que envolve a descoberta de informações em texto, em especial, entidades que podem ou não estarem catalogadas em bases de dados. A sua principal tarefa é extrair partes do texto e associar a atributos específicos (como por exemplo, nomes de pessoas, cidades, etc. (Bastos, 2006)) para o entendimento da linguagem natural.

Hotho (Hotho, Nürnberger, & Paaß, 2005) argumenta que a tarefa de IE pode ser decomposta numa série de passos de processamento, tipicamente, incluindo: itemização, segmentação de frases, atribuição gramatical e identificação de nomes de entidades (como por exemplo: nome de pessoa, nomes de locais, nomes de organizações.).

2.3.7 Pós-processamento do texto

Nas fases anteriores do processamento de texto, as técnicas de descoberta de conhecimento em textos geram usualmente formas intermédias, resultantes das análises de corpus, preparadas para um pós-processamento. Essas formas intermédias encontram-se tradicionalmente organizadas em estruturas de dados tipo matriz ($Atributo, Valor$), onde os atributos e valores servem de base aos diversos processos subsequentes, tais como agrupamento, indexação, visualização de informações, etc..

Segundo Gershon (Gershon, 1977), a visualização de informação deve socorrer-se de mecanismos que permitam melhorar a sua perceção, como por exemplo aspetos de computação gráfica, HCI (*Human-Computer Interaction* ou Interfaces Homem-Máquina) ou mineração de dados, minimizando desta forma alguns dos problemas em lidar com o enorme volume de dados disponíveis. De acordo com Freitas e outros (Freitas, Chubachi, Luzzardi, & Cava, 2001), a representação visual da informação deverá abstrair os detalhes do conjunto de informações para a organizar segundo um critério e técnica de visualização adequado.

Importa assim na fase de pós-processamento do texto escolher a forma mais correta de visualização da informação extraída e, de acordo com a tipologia da informação a apresentar, permitindo por um lado uma correta e rápida apreensão da informação, mas por outro cingindo a visualização aos aspetos cruciais e fulcrais da análise pretendida.

3 Metodologia de Investigação

Num processo de investigação importa definir claramente os princípios metodológicos e respetivos métodos a utilizar. No contexto deste trabalho a investigação traduziu-se essencialmente pela pesquisa nas áreas de Web Semântica (foco inicial pretendido) e de boas práticas na ITIL, culminando com a união destas duas componentes pelo desenho e teste de um modelo semântico para suporte a uma função de ITIL. Para tal, foi necessário recorrer a diferentes metodologias de investigação, adequadas às diferentes etapas da elaboração, desde a recolha de informação e bases de suporte às decisões tomadas, até à conceção e aplicação do modelo a um caso real.

Explicam-se de seguida algumas metodologias e estratégias de investigação usadas.

3.1 Metodologias de Investigação

A ação de investigação envolve não só a pesquisa de informação, mas também um conjunto de métodos e técnicas que permitam uma abordagem eficaz a observação pretendida. Neste prisma, e segundo Saunders e outros (Saunders, Lewis, & Thornhill, 2003), o investigador é independente dos dados observados e não tem influência sobre os mesmos, relevando uma corrente de **Positivismo** na abordagem à investigação. Segundo o mesmo autor, a abordagem à luz do **Interpretativismo**, compreende um investigador que não só se insere no contexto de investigação como também pesquisa o significado dos dados no sentido de lhes atribuir coerência. Já para o **Realismo**, existem interpretações partilhadas e que são independentes da subjetividade humana.

Também segundo os referidos autores, o método dedutivo é adequado para explicar relações entre variáveis qualitativas ou quantitativas, pressupondo que o investigador é independente dos dados que está a observar.

O método indutivo, segundo Saunders e outros (Saunders, Lewis, & Thornhill, 2003), estabelece que a realidade não é redutível a um conjunto de variáveis, mas antes pelo contrário, cada caso deve ser observado no seu meio envolvente e analisado nessa circunstância.

3.2 Estratégias de Investigação

O processo de elaboração da dissertação assentou em diferentes estratégias de investigação, de acordo com os objetivos específicos a atingir em cada etapa. Muito embora se identificar prioritariamente enquanto um Estudo de Caso, outras estratégias foram utilizadas na recolha e tratamento de dados e informação, nomeadamente a estratégias de *Grounded Theory* e *Action Research*.

3.2.1 Estudo de Caso

O estudo de caso tem sido uma das ferramentas utilizadas por vários investigadores e em contextos diferentes. As principais linhas orientadoras segundo Yin (Yin, 1994), Merriam (Merriam, 1998), Stake (Stake, 1995), Miles e Huberman (Miles & Huberman, 1994), Gillham (Gillham, 2001) são:

- O estudo de caso deve estar focado num objeto de estudo;
- O estudo de caso deve ser real;
- O estudo de caso deve possuir um contexto específico.

De acordo com Merriam (Stake, 1995), um estudo de caso deve ter também como aspeto principal o objeto de estudo, enquanto o método de investigação é secundário.

Yin (Yin, 1994) salienta o interesse na abordagem metodológica do estudo de caso. Este autor define o estudo de caso como uma investigação empírica que se destina a "*[..] investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomena and context are not clearly evident"*.

O estudo de caso deve ter um "caso" que é o objeto do estudo. O "caso" deve ser uma unidade funcional e organizacional complexa, ser investigado no seu contexto natural através de métodos múltiplos e também ser contemporâneo. No entanto, os investigadores mencionados enfatizam diferentes características.

Stake (Stake, 1995) salienta que a questão fundamental num caso de estudo não é o método da investigação, mas sim o objeto do estudo.

3.2.2 Grounded Theory

A *Grounded Theory* é uma das metodologias qualitativas de investigação quantitativa que tem vindo progressivamente a ser mais utilizada pelos investigadores no âmbito das ciências sociais e humanas, nomeadamente em diferentes áreas da psicologia.

Os desenvolvimentos mais recentes da investigação qualitativa tendem a adaptar uma posição epistemológica não positivista, recorrendo a procedimentos metodológicos que envolvem uma análise mais detalhada e flexível de material escrito, verbal ou visual, que não é convertido em pontos ou escalas numéricas, nem é considerado um espelho de uma realidade externa objetiva (Seale, 2000). Este método não procura encontrar modelos abstratos de conhecimento, sendo particularmente utilizado para a compreensão das experiências e dos significados que os seres humanos constroem na sua interação.

Por isso, a metodologia qualitativa é utilizada em estudos que contextualizam o conhecimento, tomando o próprio processo de construção de conhecimento como uma dimensão importante a considerar. Este posicionamento suporta-se na crença de que não existe produção de conhecimento independente do sujeito conhecedor, assumindo-se que o investigador deve incorporar e assumir na sua produção científica a sua própria subjetividade.

Segundo Saunders e outros (Saunders, Lewis, & Thornhill, 2003), é um método composto por uma aproximação indutiva e dedutiva. A recolha de dados inicia-se sem a formulação de uma estratégia de investigação. A teoria é desenvolvida á medida que os dados se vão interpretando. Cada amostra de dados pressupõe uma teoria de inferência sobre a amostra seguinte. A investigação pára quando já não existe mais valor acrescentado à teoria.

3.2.3 Action Research

O *Action Research* é uma metodologia quantitativa que se baseia na introdução de alterações na organização, avaliar os resultados das mesmas e atuar de seguida numa espiral formada por análise-ação-avaliação. Segundo Saunders e outros (Saunders,

Lewis, & Thornhill, 2003), o investigador é parte da organização em que a investigação se está a desenvolver.

Tal como descrito por Eden e Huxham (Eden & Huxham, 1996), a *Action Research* assenta nas seguintes caraterísticas:

- utilização de um processo iterativo de identificação do problema, planeamento, ação e avaliação, que conduz a avanços teóricos em pequenos passos incrementais;
- os resultados da investigação devem ter implicações teóricas que vão para além da resolução do problema concreto;
- a investigação deve conduzir a generalizações que possam ser expressas através de ferramentas, técnicas, modelos e métodos aplicáveis noutras situações.

3.2.4 Secondary Data

No âmbito da pesquisa e análise de dados podem-se utilizar dados já existentes que podem ser registos provenientes de base de dados, relatórios, publicações diversas, atas e sumários de reuniões, entre outros, (Saunders, Lewis, & Thornhill, 2003). Estes autores defendem ainda que para alguns tipos de projetos, *secondary data* poderá ser a principal fonte de informação para responder às questões e objetivos da investigação. A *secondary data* pode ser qualitativa ou quantitativa. No contexto do presente estudo existem vários tipos de dados para os quais os referidos autores identificam as seguintes as fontes:

- Normas e documentos publicados;
- Protocolos de boas práticas;
- Estudos efetuados;
- Dados existentes nos sistemas de informação.

Segundo Andersen, Prause, e Silver (Andersen, Prause, & Silver, 2011) uma das vantagens da utilização do *secondary data* reside no facto de num conjunto grande de dados ser possível agregar variáveis e criar ficheiros apropriados para as diferentes questões de investigação ou métodos estatísticos, sendo apontada como principal

desvantagem a limitação na escolha de questões específicas ou métricas na recolha dos dados.

Na realidade, o investigador está limitado às metodologias utilizadas no processo de aquisição dos dados. A sua reutilização depende dos métodos utilizados aquando da investigação original.

3.3 Métodos de Investigação

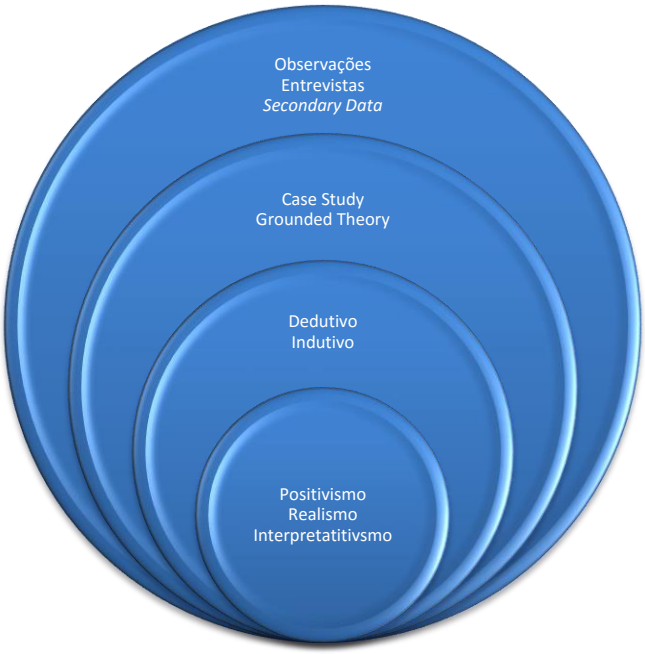
3.3.1 Métodos Qualitativos e Quantitativos

Segundo Creswell (Creswell, 2003), um método qualitativo é aquele em que o investigador utiliza estratégias de narrativas, fenomenológicas, etnográficas e *grounded theory* para desenvolver o conhecimento sobre os dados. Em alternativa, no caso de o investigador utilizar métricas que lhe permitam efetuar análises estatísticas baseadas em dados numéricos, estamos perante a utilização de um método de investigação quantitativo.

3.3.2 Justificação da Abordagem Metodológica de Estudo de Caso

A abordagem metodológica escolhida para a investigação nesta dissertação, assenta maioritariamente e na sua essência num estudo de caso, recorrendo à utilização de multimétodos que, segundo Creswel (Creswell, 2003) consistem na orientação do estudo segundo o conhecimento existente e na orientação da pesquisa para o problema, onde as estratégias escolhidas envolvem a recolha de dados de diferentes tipos, quantitativos ou qualitativos.

Ilustração 9 - Métodos de investigação



Fonte: Adaptado de Saunders e outros (Saunders, Lewis, & Thornhill, 2003)

Assim, atendendo a que na presente dissertação existem normas e teorias investigadas com recurso a métodos interpretativistas e indutivos, e dados e informações recolhidos e trabalhados com base em métodos dedutivos e positivistas, a abordagem escolhida assenta numa metodologia multimétodo.

4 Desenvolvimento do Modelo de Inteligência Semântica

Este capítulo debruça-se sobre a construção e desenvolvimento do modelo de inteligência semântica, salientando aspetos relativos à recolha de informação inerente à função da ITIL em estudo, nomeadamente o "*Service Desk*" (apoio ao utilizador). Assim, o estudo começa por identificar as principais e mais importantes áreas (serviços de TI) onde o apoio informático é como se irá constatar identificado, por parte dos utilizadores (gestores, consumidores de recursos TI, etc.), como mais crítico. De seguida é aferida a perceção dos utilizadores sobre qual seria, sob a perspetiva de melhoria do serviço de suporte informático, a medida com melhores resultados esperados de entre um conjunto de medidas, interpretados os respetivos resultados, e obtidas algumas conclusões relevantes para as ações a tomar no desenho e criação do modelo, orientando-o assim para dar resposta às necessidades identificadas.

4.1 Identificação dos Perfis de Serviços de TI Relevantes

O "*Service Desk*" encontra-se inserido no processo de "*Service Operation*" da ITIL, tal como descrito em 2.1.3.4 *Service Operation* (Operação do Serviço) e existe para garantir um ponto único de contacto (*Single Point of Contact* - SPOC) entre os utilizadores de recursos de TI e o *Service Management*. Utiliza as boas práticas da ITIL para gerir todo e qualquer fluxo de informação relacionado com as TI, em alinhamento com a estratégia da instituição.

Assim, toda e qualquer solicitação relacionada com recursos de TI será tratada pelo *Service Desk*, o que por si, e atendendo à natural evolução e dependência crescente maior destas tecnologias, obriga a que este tenha de gerir um cada vez maior volume de informação, de tecnologias e de recursos distintos, cada vez mais diversificados. Identificam-se de seguida alguns recursos de TI frequentemente disponibilizados em instituições (Andrew & Rob, 2014):

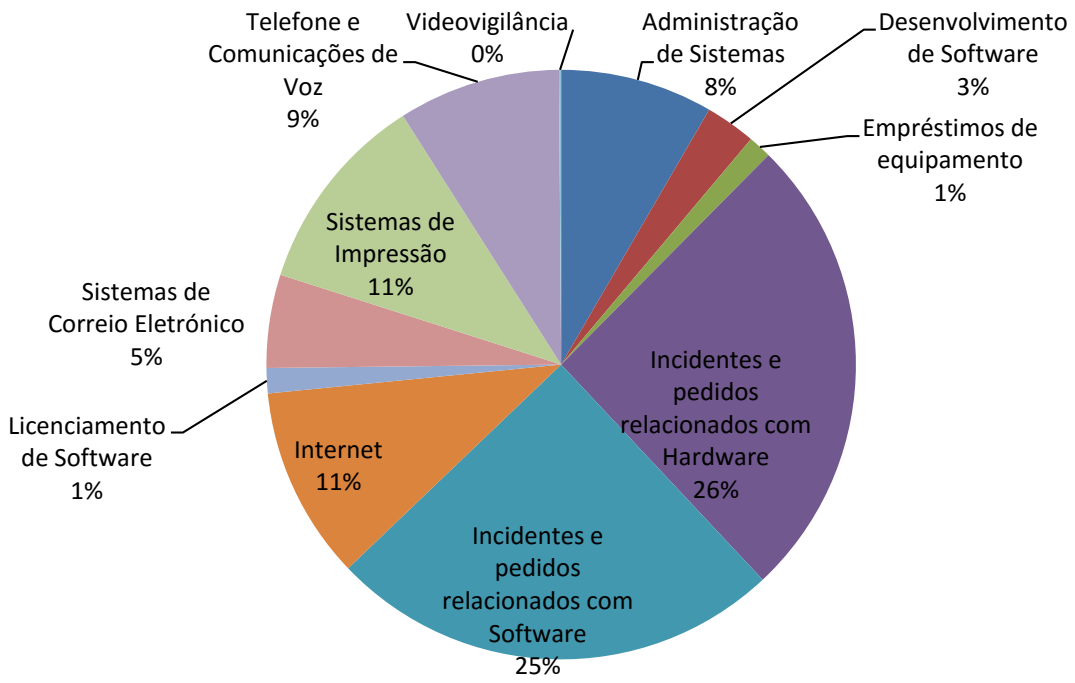
- **Administração de Redes**: gestão de equipamentos ativos de rede e recursos relacionados com mecanismos de comunicação de dados;
- **Administração de Sistemas**: gestão de aplicações e sistemas operativos servidor, bem como os demais serviços cujo funcionamento seja crucial para

outros sistemas e aplicações, como por exemplo, bases de dados, permissões e diretório corporativo, politicas de gestão centralizadas, etc.

- **Desenvolvimento de Software**: desenvolvimento de aplicações de suporte ao negócio;
- **Empréstimos de equipamento**: gestão de inventário e equipamentos;
- **Incidentes e pedidos relacionados com Hardware**: pedidos de mudança (alteração) ou reporte de incidentes relacionados com *hardware*;
- **Incidentes e pedidos relacionados com Software**: pedidos de mudança (alteração) ou reporte de incidentes relacionados com *software*;
- **Internet**: mecanismos e equipamentos relacionados com os acessos à Internet, desde controlo de larguras de banda, conteúdos e acessos
- **Licenciamento de Software**: pedidos de aquisição de contratos para utilização de *software*;
- **Sistemas de Correio Eletrónico**: sistemas que permitem a gestão de correio eletrónico, agendas e funcionalidades acessórias relacionadas com a gestão de tempo;
- **Sistemas de Impressão**: gestão pedidos de mudança ou reporte de incidentes relacionados com equipamentos e recursos de impressão/digitalização e respetivo controlo, como impressoras, *scanners*, faxes, etc.
- **Telefone e Comunicações de Voz**: gestão de equipamentos terminais de comunicações de voz (telefones, telemóveis, centrais de alarme, etc.) e equipamentos centrais relacionados com mecanismos de comunicação de voz;
- **Videovigilância**: gestão de equipamentos terminais de videovigilância (câmaras, gravadores, cablagem, etc.) e equipamentos centrais relacionados com mecanismos de videovigilância.

No gráfico 1 é possível observar a distribuição por número de incidentes pelas principais categorias identificadas, denotando-se uma clara prevalência das categorias mais genéricas, nomeadamente “Incidentes e pedidos relacionados com Hardware” e “Incidentes e pedidos relacionados com Software”.

Gráfico 1 - Distribuição de incidentes por principais categorias



Adaptado de (Andrew & Rob, 2014)

A interpretação realizada pelo autor dos dados deste gráfico do estudo é somente na perspetiva de número de incidentes, não tendo existido uma preocupação com a alocação do tempo consumido ou mesmo custos indexados na resolução dos incidentes, isto porque a resolução de um problema mais especializado ou específico, embora com uma duração inferior, poderá ter um custo de resolução superior.

Desta forma, e para uma melhor quantificação da relevância das categorias onde focar a presente dissertação, houve necessidade de elaborar um questionário direcionado para diferentes segmentos de utilizadores, e que pretende aferir, entre outras questões, quais os serviços ou categorias de incidentes e pedidos tidos como mais importantes ou cruciais na utilização diária dos recursos das TI, bem como quais as ações a tomar por forma a melhorar os serviços de suporte das TI. Desta forma, e tomando como partida as categorias identificadas pelo estudo foi elaborado o questionário em anexo (**anexo 1**), cujos resultados e respetiva análise será apresentada nos pontos seguintes.

4.1.1 Impacto da Aplicação da ITIL nas Organizações

Tal como descrito no capítulo 2.2.1, Gestão de Mudança na ITIL, a implementação de boas práticas baseadas na ITIL pode levar à otimização e alteração de procedimentos organizacionais. Implementar uma mudança numa organização pressupõe que dessa mudança irão resultar ganhos de performance (de tempo, financeira, produtividade, etc.). Desta forma, e apesar de as vantagens de adoção das boas práticas ITIL estarem patentes em praticamente toda a documentação e casos de sucesso, importa aferir a motivação e impacto previsto destas medidas. Algumas das vantagens são referidas por Zisblat (Zisblat, 2008), como por exemplo:

- Redução nos tempos médios de resposta a solicitações;
- Redução nos tempos médios de resolução de incidentes;
- Garantia de maior controlo nas alterações ao ambiente;
- Redução da perda de produtividade causada pela indisponibilidade de recursos;
- Aumento da satisfação global dos utilizadores;
- Aumento da satisfação global dos clientes.

De acordo com o estudo da Ayehu (Ayehu, 2015) as tarefas de suporte a incidentes podem ser também automatizadas, garantindo assim:

- Uma melhor preparação para eventuais repetições do mesmo incidente;
- Um registo de informação mais rico e preciso;
- Uma resposta vinte e quatro horas por dia, sete dias por semana;
- Que o erro é reduzido e são assegurados registos assertivos;
- Que o problema seja contido rapidamente, isto é, que seja sanado antes de tomar proporções maiores.

Desta forma, a implementação de uma abordagem com vista à otimização ou automatização de processo de suporte das TI estima-se que se traduza numa melhoria dos custos financeiros, alocação de recursos humanos, e no global de melhoria da prestação de toda a organização.

4.1.2 Análise das Respostas aos Questionários

O questionário em anexo (**anexo 1**) foi construído com o intuito de endereçar lacunas na informação pesquisada, nomeadamente quais os serviços de TI mais críticos, do ponto de vista do utilizador, quais as medidas a implementar com vista à melhoria da qualidade no apoio ao utilizador, encontrando-se as respostas alocadas por tipo ou natureza de atividade da organização.

O questionário foi endereçado a quarenta e três contactos, desde gestores e técnicos responsáveis por gerir estruturas de TI, a utilizadores de recursos de TI em diversas organizações (empresas – 14 contactos, instituições sem fins comerciais – 13 contactos e da administração pública – 16 contactos), tendo sido recebidas vinte e respostas, o que perfaz um rácio de perguntas respondidas de 58%.

Tabela 6 - Número de pedidos de preenchimento vs número de respostas ao questionário

Número de pedidos de preenchimento	43
Número de respostas obtidas	25
Rácio	**58%**

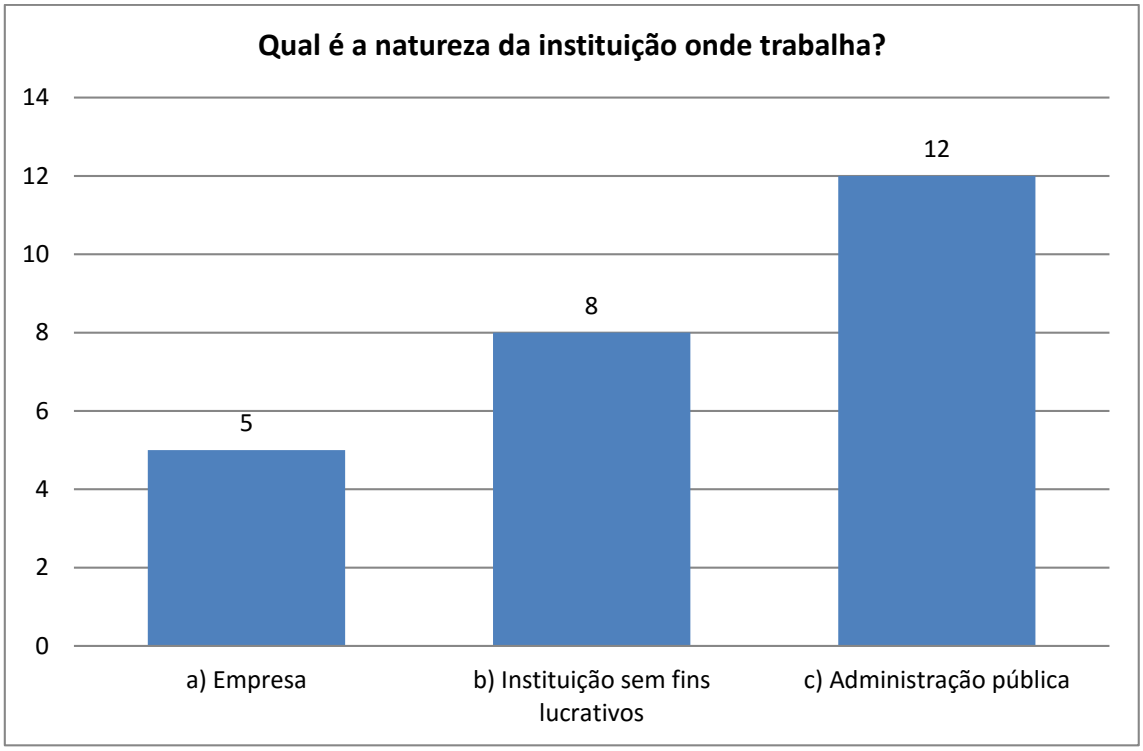
Fonte própria

A primeira questão endereçada foi "*Qual é a natureza da instituição onde trabalha?*".

Esta questão pretende enquadrar as respostas consoante a tipologia das organizações, podendo ser relevante aferir a existência de padrões que sejam ocasionados pelo facto de a natureza das organizações ser divergente.

As respostas obtidas para esta questão podem ser observadas no gráfico 2.

Gráfico 2 - Respostas ao questionário (questão 1)



Fonte própria

Analisando as respostas (gráfico 2 e tabela 7), entende-se que a apesar da distribuição dos pedidos de resposta endereçados ter sido em termos de igualdade de número por organização semelhante (33%, 30% e 37% respetivamente para "Empresa", "Instituição sem fins lucrativos" e "Administração Pública" respetivamente), as respostas tiveram uma aderência mais forte na componente de "Administração Pública", com 48% do total das respostas, de 32% na categoria de "Instituição sem fins Lucrativos" e 20% para "Empresa".

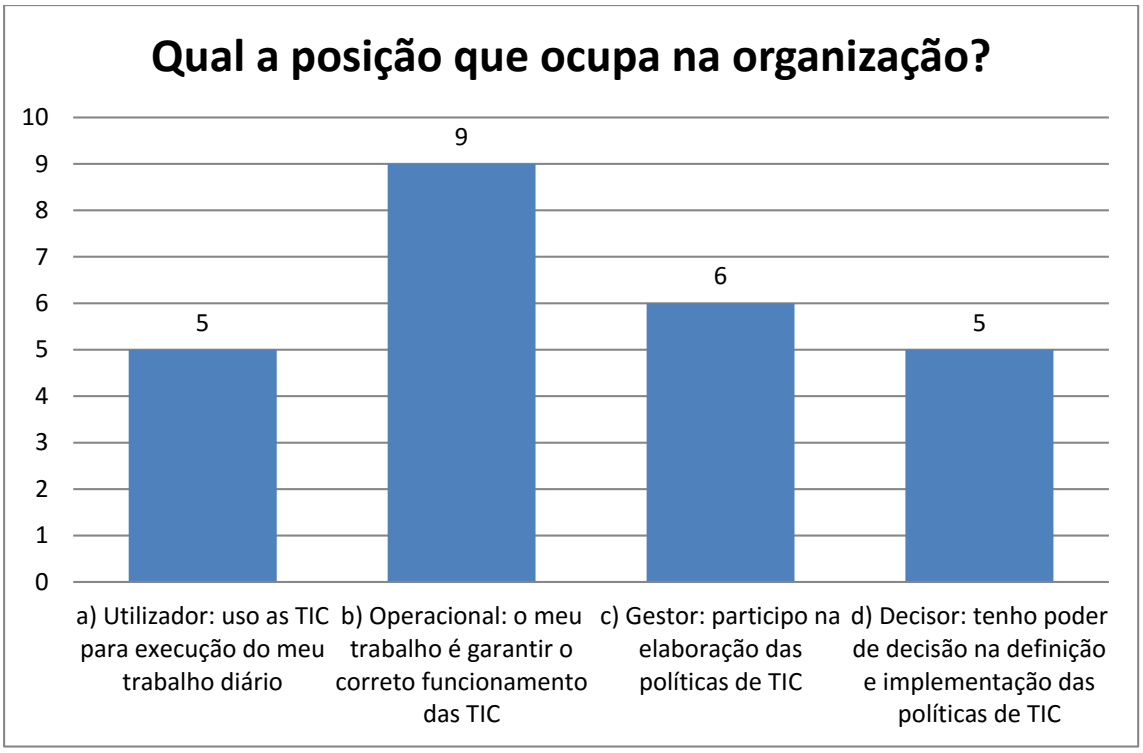
Tabela 7 - Número de questionários respondidos por tipo de organização

Tipo de organização	Pedidos de preenchimento	Respostas
Empresa	14	5
Instituição sem fins lucrativos	13	8
Administração Pública	16	12
Total	**43**	**25**

Fonte própria

A segunda questão pretende estabelecer paralelismos entre a posição ocupada pela pessoa que responde a escolha das prioridades de criticidade dos serviços de TI.

Gráfico 3 - Respostas ao questionário (questão 2)

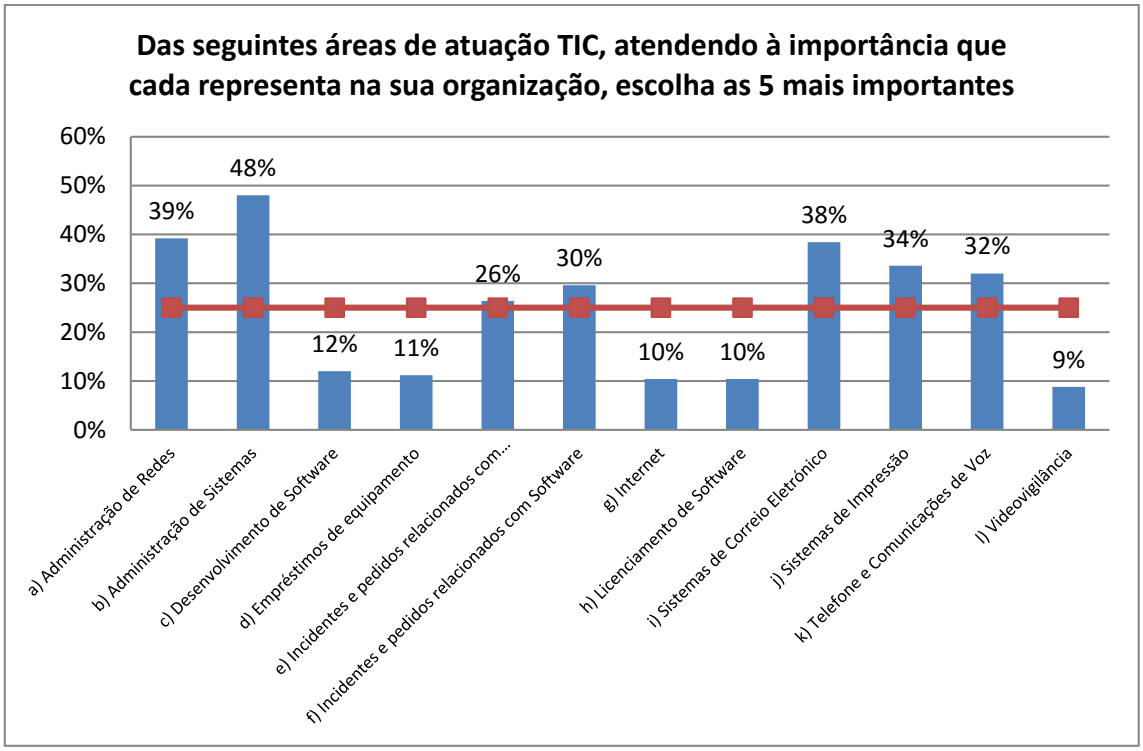


Fonte própria

Nesta resposta os inquiridos pertencem na sua maioria às esferas operacionais, encontrando-se as restantes categorias distribuídas quase equitativamente. Não foi estabelecido um paralelismo ou padrão de resposta à questão 2 em função da posição dos participantes no inquérito, pelo que se concluí que a função não influenciou, neste inquérito, a visão sobre as prioridades dos recursos de TI.

Tal como descrito no capítulo 4.1, existiu a necessidade de clarificar e cingir o estudo às áreas de atuação ou recursos de TIC considerados mais críticos ou importantes. Para tal, a segunda questão elaborada foi "*Das seguintes áreas de atuação TIC, atendendo à importância que cada representa na sua organização*" e permitia na resposta a priorização das 5 áreas de atuação de TIC mais importantes, com os resultados descritos no gráfico 4.

Gráfico 4 - Respostas ao questionário (questão 3)



Fonte própria

Para o cálculo da classificação de cada categoria foi tida por base a ordenação atribuída em cada resposta, tendo sido feita uma contagem de frequências de posição escolhida para cada categoria, que se apresenta na tabela 8.

Tabela 8 – Respostas ao questionário - frequências absolutas (questão 3)

Ordenação Escolhida	a) Administração de Redes	b) Administração de Sistemas	c) Desenvolvimento de Software	d) Empréstimos de equipamento	e) Incidentes e pedidos relacionados com Hardware	f) Incidentes e pedidos relacionados com Software	g) Internet	h) Licenciamento de Software	i) Sistemas de Correio Eletrónico	j) Sistemas de Impressão	k) Telefone e Comunicações de Voz	l) Videovigilância
5	3	4	0	0	2	4	0	0	5	4	3	0
4	4	5	0	1	2	2	0	1	4	3	3	0
3	3	4	1	1	4	0	3	1	1	2	3	2
2	2	3	3	3	1	4	1	3	1	1	1	2
1	5	2	6	1	1	1	2	0	2	2	2	1

Fonte própria

Posteriormente, a frequência absoluta foi multiplicada pelo número da ordem respetiva, tendo sido obtido um valor global por categoria entre 0 e 25 por cada resposta (ex.: uma categoria escolhida em 3 respostas como primeira prioridade sem nenhuma outra escolha nas restantes respostas, terá uma classificação de 15 pontos – 5 x 3), e para o universo das 25 perguntas, entre 0 e 125, valores esses que foram usados para apurar as frequências relativas respetivas (tabela 9).

Tabela 9 - Respostas ao questionário - frequências relativas (questão 3)

	a) Administração de Redes	b) Administração de Sistemas	c) Desenvolvimento de Software	d) Empréstimos de equipamento	e) Incidentes e pedidos relacionados com	f) Incidentes e pedidos relacionados com	g) Internet	h) Licenciamento de Software	i) Sistemas de Correio Eletrónico	j) Sistemas de Impressão	k) Telefone e Comunicacões de Voz	l) Videovigilância
5	3	4	0	0	2	4	0	0	5	4	3	0
4	4	5	0	1	2	2	0	1	4	3	3	0
3	3	4	1	1	4	0	3	1	1	2	3	2
2	2	3	3	3	1	4	1	3	1	1	1	2
1	5	2	6	1	1	1	2	0	2	2	2	1
Class. Abs. (escala 0 a 125)	49	60	15	14	33	37	13	13	48	42	40	11
Class.abs. %	39%	48%	12%	11%	26%	30%	10%	10%	38%	34%	32%	9%

Fonte própria

Observando o gráfico encontra-se um claro destaque na prioridade para a categoria "Administração de Sistemas", liderando com 48% das respostas enquanto 1º lugar, e com uma classificação final de 9% a categoria "Videovigilância". Para decisão da escolha das categorias mais críticas, usou-se a média aritmética, descartando-se categorias que tivessem classificações abaixo da média (assinalado no gráfico 4 com a linha vermelha), das quais resultaram as categorias (com classificação relativa) indicadas na tabela 10.

Tabela 10 - Respostas ao questionário - Ordenação das categorias acima da média (questão 3)

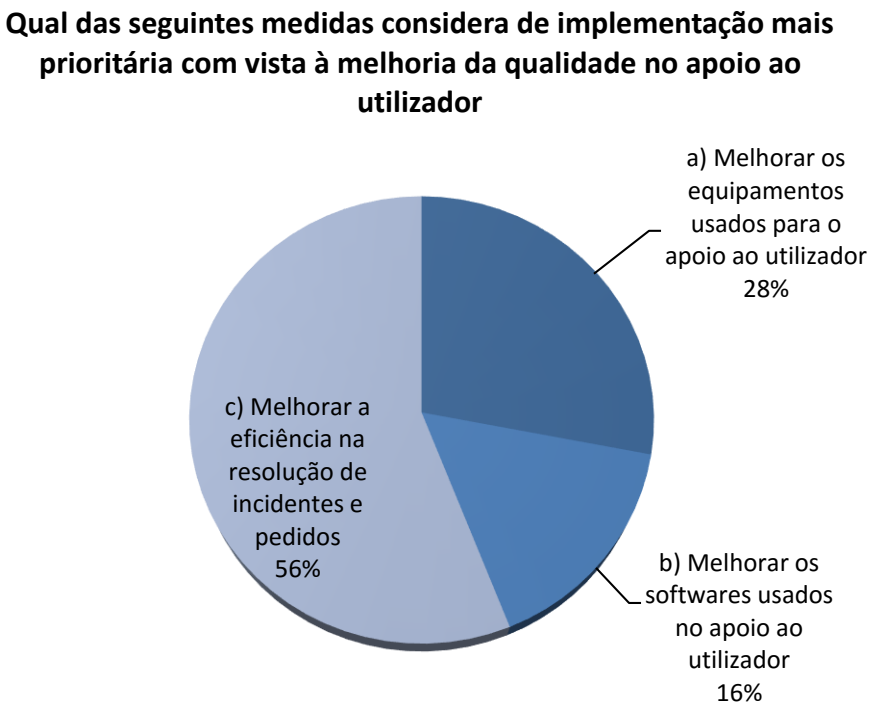
Ordenação das categorias acima da média	Classificação relativa
b) Administração de Sistemas	19%
i) Sistemas de Correio Eletrónico	16%
a) Administração de Redes	16%
j) Sistemas de Impressão	14%
k) Telefone e Comunicações de Voz	13%
f) Incidentes e pedidos relacionados com Software	12%
e) Incidentes e pedidos relacionados com Hardware	11%

Fonte própria

Assim, o estudo irá debruçar-se sobre as sete categorias de serviços TI resultantes, consideradas com mais importantes pelas organizações (pessoas) inquiridas.

Por último, e com o intuito de identificar prioridades na conceção do modelo de para que o mesmo as possa endereçar proficuamente, surge a questão "*Qual das seguintes medidas considera de implementação mais prioritária com vista à melhoria da qualidade no apoio ao utilizador*", apresentando-se as respostas no gráfico 5.

Gráfico 5 - Respostas ao questionário (questão 4)



Fonte própria

Da análise do gráfico, percebe-se que a esmagadora maioria dos inquiridos entende como a principal medida para melhorar a qualidade dos serviços de TI e apoio ao utilizador passa por "*Melhorar a eficiência na resolução de incidentes e pedidos*". Ora esta melhoria poderá traduzir-se numa intervenção no processo de "*Service Desk*" que resulte em ganhos operacionais, como por exemplo a automatização de parte do mesmo recorrendo a tecnologias de inteligência semântica.

4.2 Pressupostos do Modelo a Conceber

Analisados que estão alguns dos pressupostos, importa iniciar o esboço do modelo e definir em traços gerais os pressupostos onde o modelo irá assentar. Assim, e de acordo com as conclusões obtidas no ponto anterior, o modelo deverá obedecer às regras base de um processo, transformando *inputs* em *outputs*, e terá como principal objetivo o de melhorar, automatizar o otimizar a gestão do "*Service Desk*" numa organização, tendo como resultados esperados uma melhoria no serviço em termos de eficiência. Estando também identificadas as categorias de serviços mais críticos, entende-se que os

processos a melhorar tenderão a estar relacionados com estas categorias, pelo que o modelo a conceber, embora possa ser adaptado genericamente a diversos tipos de serviços de TI será, neste estudo, indexado às categorias identificadas.

Ilustração 10 - Fluxo de um pedido ou incidente

Fonte própria

Para uma melhor afetação dos recursos existentes na resolução de solicitações é importante uma correta categorização das mesmas, dado que desta forma torna-se possível que uma aplicação de gestão de solicitações proceda à automática atribuição e respetivo escalonamento das equipas mais adequadas, otimizando desta forma recursos humanos no processo de suporte ao utilizador (ilustração 10).

Estima-se que esses processos automatizados revejam com a inteligência de análise léxica os factos, obviando erros introduzidas pelo fator humano como por exemplo a falta de conhecimento técnicos dos envolvidos no processo, o elevado número de pedidos a triar, bem como a necessidade de ler e interpretar cada solicitação o que se reflete em consumo de tempo.

Uma potencial melhoria no processo de gestão de pedidos e incidentes será o de **garantir uma triagem e categorização automática dos pedidos**, permitindo que um pedido ou incidente registado possa fluir automaticamente para a equipa específica de suporte sem necessidade de qualquer processo manual. Esta categorização automática traduz-se nas seguintes vantagens:

- Redução de erros na classificação de pedidos (triagem): em caso de incerteza, o modelo irá aguardar pela triagem manual ("aprendendo" com este processo), sendo que nos restantes casos (com algum grau de certeza) a triagem é imediata;
- O tratamento das intervenções é mais célere, já que podem fluir diretamente do utilizador para as equipas de suporte;
- Aprendizagem automática, que permitirá adaptar-se rapidamente a novas tecnologias e alterações à organização;

Para a recolha de *inputs*, é necessário que existam fontes de dados, pelo que este modelo irá pressupor a existência de um sistema informático onde assentará o SPOC, e serão registadas e classificadas todas as ocorrências (incidentes e pedidos). O serviço de apoio ao utilizador deverá então assentar o seu funcionamento nas boas práticas da ITIL no que concerne ao "*Service Desk*".

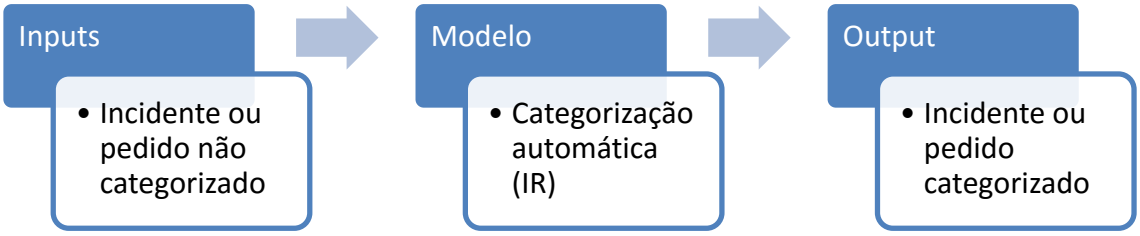
Para a implementação do modelo, é necessário que a organização tenha em atenção as seguintes pré-condições:

- **Conhecimento da organização da norma ITIL**, especialmente nos serviços de suporte de TI, garantindo uma linguagem coerente entre os diversos elementos de suporte;

- **Existência de SPOC implementando e difundido**: todos os incidentes e pedidos devem ser registados;
- **Registo e categorização de incidentes e pedidos em sistema informático**, permitindo o seu acompanhamento, triagem e posterior encaminhamento para as equipas de suporte corretas;
- **A categorização de incidentes e pedidos deverá obedecer a um conjunto predefinido de categorias**, com uma relação direta ou quase direta para as equipas de suporte. Ex.: um incidente ou pedido categorizado como (incidente em…) "Correio Eletrónico", deverá ser atribuído para resolução às equipas que lidam com estas tecnologias.

Cumpridas estas pré-condições, torna-se possível aplicar o modelo de classificação automática de pedidos e incidentes (ilustração 11).

Ilustração 11 - Diagrama conceptual do modelo



Fonte própria

Em maior detalhe, o modelo terá como *input* incidentes e pedidos (intervenções) não categorizados, sobre os quais fará um processamento com base a metodologias de inteligência semântica (em maior detalhe no capítulo 4.3), dos quais resultará a atribuição de uma categoria que ficará associada ao registo da intervenção respetiva. Pressupõe-se que este processo apresente uma redução significativa no tempo necessário à triagem e tratamento inicial dos pedidos.

4.3 Processos de Inteligência Semântica Aplicáveis

A prospeção de dados é formada por um conjunto de ferramentas e técnicas que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neuronais e

estatística, são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. Este conhecimento pode ser apresentado por essas ferramentas de diversas formas: agrupamentos, hipóteses, regras, árvores de decisão, grafos, ou dendrogramas.

Neste capítulo pretende-se distinguir dos processos e metodologias identificados no capítulo 2.3 quais os mais adaptados ao modelo a conceber, encontrando os mais adaptados para extração de informação sobre linguagem textual.

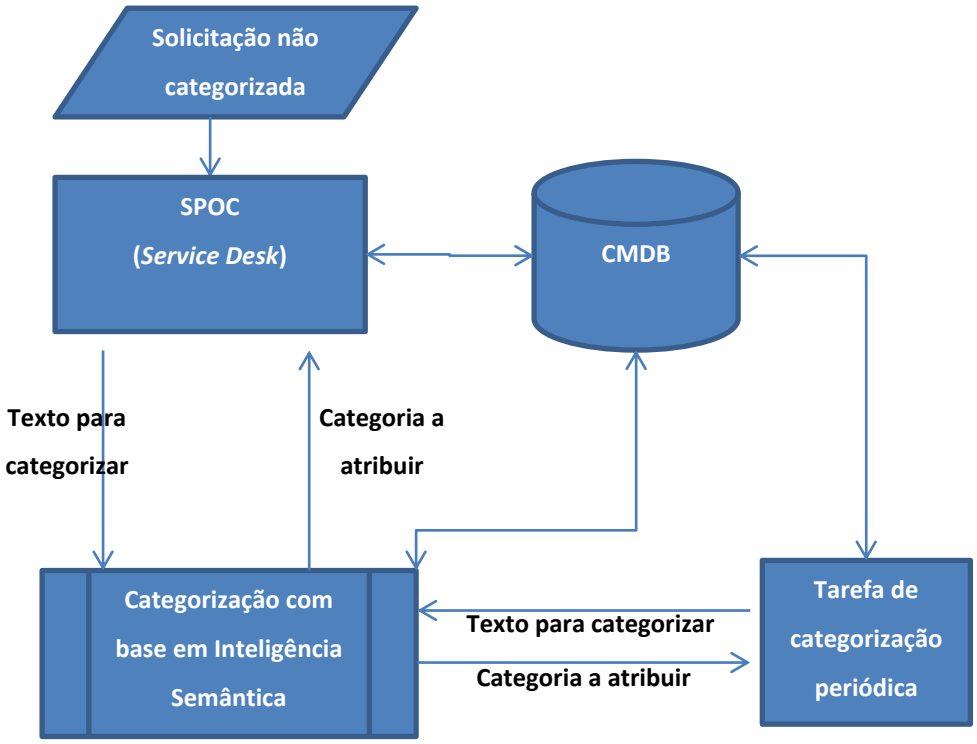
4.3.1 *Inputs, Outputs* e Processos do Modelo

Definido que está que os *inputs* do modelo serão campos de linguagem textual de um sistema informático onde assenta o "*Service Desk*", a recolha desses dados obriga à existência de conetores que permitam tecnicamente este acesso à informação. Em maior detalhe técnico (ilustração 12), essa abordagem poderá passar por obter um acesso direto à base de dados respetiva, criação de *webservices*[5] ou consumo de *webservices* já disponíveis que permitam as operações de consulta (e posteriormente escrita), podendo a tarefa de categorização de pedidos ser efetuada a pedido (por exemplo, sempre que se altera o texto com o descritivo do pedido ou surge um pedido novo) ou executada periodicamente (sobre pedidos ainda não triados).

O modelo poderá funcionar de forma conjunta com o sistema de informação de suporte ao "*Service Desk*", ou de forma autónoma, servindo apenas enquanto elemento função de consulta (por exemplo, disponível através de *webservice*), podendo neste caso ser extrapolado e utilizado noutras aplicações e sistemas.

[5] *Webservice* é uma solução utilizada na integração de sistemas e na comunicação entre aplicações diferentes. Com esta tecnologia é possível que novas aplicações possam interagir com aquelas que já existem e que sistemas desenvolvidos em plataformas diferentes sejam compatíveis.

Ilustração 12 - Arquitetura Técnica de Funcionamento do Modelo



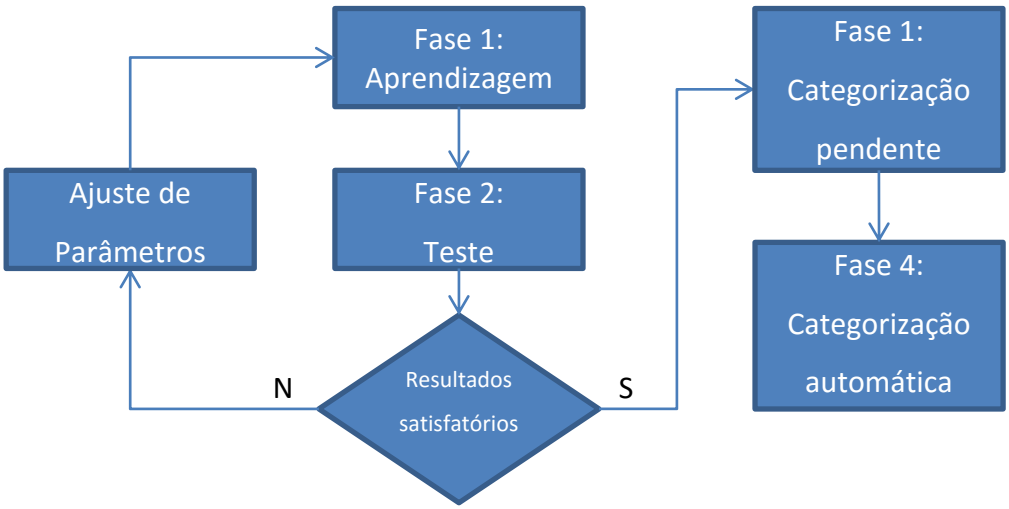
Fonte própria

Como em todos os processos de mudança, é necessário aferir e controlar os resultados obtidos e compará-los com os resultados esperados. Assim, a implementação deste modelo deverá obedecer às seguintes fases lógicas (ver também a ilustração 13):

- **Fase 1**: Aprendizagem com recurso a casos já categorizados;
- **Fase 2**: Teste do modelo com um conjunto de casos de teste; caso os valores obtidos não sejam satisfatórios, poderão ser ajustados parâmetros do modelo por forma a otimizar os resultados;
- **Fase 3**: Em categorização pendente, isto é, a categorização é assegurada pelo sistema de inteligência semântico, mas qualquer outra ação que dependa desta categorização não é executada até que a categoria seja confirmada manualmente pela equipa de triagem;
- **Fase 4**: Em categorização automática. O sistema é totalmente autónomo. Assume-se que as categorizações que o sistema faz são fidedignas. As solicitações que o sistema

não consiga atribuir uma categoria continuarão a ser triadas manualmente, sendo o resultado usado para a aprendizagem do modelo.

Ilustração 13 - Fases de Implementação do Modelo



Fonte própria

Apesar da identificação das categorias mais importantes estar assegurada, o modelo é agnóstico à limitação das mesmas, já que a primeira fase da implementação do modelo passará pela aprendizagem com recurso a um conjunto de pedidos previamente categorizados. Desta forma, o sistema fica apto a categorizar novos textos. Apesar desta versatilidade, a alteração ou inclusão frequente de categorias poderá influenciar negativamente a performance do modelo, sendo essa influencia mais negativa caso se assista por exemplo a subdivisão de categorias existentes em novas categorias. Nestas situações, o sistema terá de reaprender com uma nova categorização, o que obriga a um esforço acrescido já que o processo terá de ser alimentado por pedidos previamente categorizados manualmente. Desta forma é aconselhável a estabilização das categorias numa fase inicial, processo esse que é até requisito para que seja possível delinear a distribuição das solicitações pelas equipas de suporte. Sem esta categorização essa distribuição não é facilitada.

Certo é que a realidade de cada organização varia de organização para organização, por diversos fatores desde a cultura organizacional, até aos colaboradores que a compõe.

Tal realidade inclui também terminologias e nomenclaturas que, embora bem difundidos e conhecidos no seio de uma organização, poderão ter pouco ou nenhum significado noutras. Um exemplo prático disso poderá ser por exemplo a referência interna que se dá a um equipamento: se numa organização forem utilizados números de inventário, e o equipamento identificado pelo mesmo, essa referência não é válida para outra organização que refira o equipamento pelo seu nome e local onde está fisicamente. Desta forma, na segunda fase de implementação do modelo, a fase de teste, alguns dos parâmetros do sistema poderão ser adaptados à realidade de cada organização. O modelo permite ajustar, para suprir situações como a descrita, uma lista de expressões baseadas em expressões regulares, que permitam identificar termos específicos de uma organização e transformá-los num termo mais genérico. Por exemplo a expressão regular para converter um número de inventário com 6 dígitos para a palavra "computador" (assumindo que se conhece que qualquer número de inventário com 6 dígitos se trata de um computador) é a seguinte: $[0-9]\{6\} -> "computador"$. Assim, todos os termos que correspondessem à expressão regular poderiam ser substituídos pela palavra "computador" o que facilitará a análise pelo sistema de inteligência semântica.

4.3.2 Metodologias de Inteligência Semântica Aplicáveis

Para a escolha da ferramenta de suporte enquanto motor do sistema de inteligência semântica, foram efetuadas pesquisas cujos requisitos assentaram nos seguintes pressupostos:

- Deve ser de utilização gratuita (Freeware) e passível de ser usada para fins comerciais;
- Deve suportar análise de documentos textuais e incluir as fases de pré-processamento de texto descritas no capítulo 2.3.5;
- Deve incluir sistemas de aprendizagem de modelos semânticos;
- Deve incluir API (*Application Program Interface*) para integração com os sistemas existentes;
- Deve ser eficiente e ter suporte técnico.

Desta prospeção foram identificadas enquanto ferramentas principais o GATE (*General Architecture for Text Engineering*) e o RapidMiner, ambas cumprindo os requisitos necessários.

https://gate.ac.uk/

https://rapidminer.com/

O GATE ou General Architecture for Text Engineering é uma *suite* de aplicações *opensource* desenvolvida na Universidade de Sheffield, que começou em 1995 e hoje é utilizada por uma vasta comunidade de cientistas, empresas, professores e alunos para as tarefas de análise de processamento de linguagem natural de todos os tipos, incluindo extração de informação em vários idiomas.

O GATE visa eliminar a necessidade de resolver problemas de engenharia comuns antes de fazer a pesquisa útil, ou reengenharia de processos antes de converter os resultados da investigação em aplicações. As principais funções do GATE são as seguintes:

- Modelação e persistência de estruturas de dados especializadas;
- Medidas, avaliação e análise comparativa;
- Anotações de visualização e edição, ontologias, análise de árvores, etc.;
- Linguagem de transdução de estados finitos para prototipagem rápida e eficiente implementação de métodos de análise de superfície (JAPE);
- Extração de instâncias de treino para *Machine Learning*;

Além das funções fundamentais, o GATE inclui componentes para tarefas de processamento de línguas naturais, por exemplo, análise (*parse*), morfologia, ferramentas de marcação (*tag*), recuperação de informação e componentes de extração de informação para vários idiomas.

O GATE *Developer* e o GATE *Embedded* utilizam um sistema de extração de informação (ANNIE), que foi adaptado para criar metadados RDF ou *Web Ontology Language* (OWL) para conteúdo não-estruturado (anotação semântica).

O RapidMiner é um *software* desenvolvido por uma empresa com o mesmo nome, que disponibiliza um ambiente integrado para *machine learning, data mining, text mining, predictive analytics* e *business analytics.*

O RapidMiner baseia-se em algoritmos e modelos de aprendizagem baseados em Weka e R (Norris, 2013), que podem ser incluídos a partir de extensões.

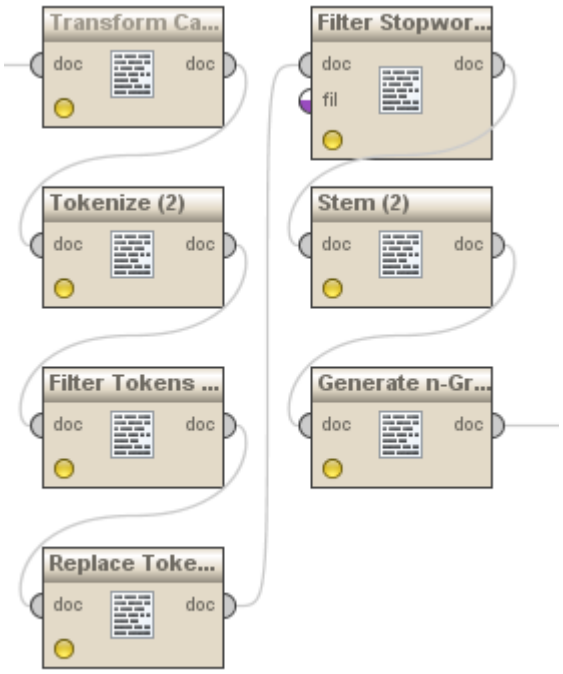
Em termos comparativos, a aplicação RapidMiner, com a sua interface gráfica (*Rapid Studio*) torna a conceção e desenho do sistema de inteligência semântico uma tarefa muito intuitiva e rápida, permitindo um acompanhamento em tempo real de todas as suas fases. Nesta área o GATE deixa bastante a desejar, oferecendo uma interface obsoleta e que obriga a alguns "truques" para que se obtenham os resultados desejados. Em termos de suporte, o RapidMiner dispõe de uma coleção de conteúdos multimédia que detalham cada operação e componente com exemplos práticos, o que facilita a aprendizagem deste sistema. A utilização do GATE implica um maior esforço na pesquisa de documentação, que assenta em textos em vez de em conteúdo visual, tornando a aprendizagem menos interessante. Por outro lado, o GATE é um sistema mais maduro, abrangendo mais idiomas que o RapidMiner (Árabe por exemplo) e que permite, tal como o RapidMiner, a instalação de extensões de funcionalidades (*plugins*), muito embora seja no GATE um processo mais complexo. Ambas as ferramentas dispõem de uma API baseada na linguagem de programação Java para integração com sistemas terceiros.

Desta forma, e dada a simplicidade de utilização e integração que o RapidMiner permite, foi esta a ferramenta escolhida para utilização no âmbito desta dissertação.

Tal como descrito no capítulo 2.3, existe um conjunto de metodologias para reconhecimento e extração de informação a partir de documentos textuais. Para que a análise seja possível, é imprescindível assegurar a passagem dos documentos textuais

por uma fase de preparação. O RapidMiner dispõe desses componentes, resultando no processo identificado na ilustração 14.

Ilustração 14 - Fase de pré-processamento de texto no RapidMiner



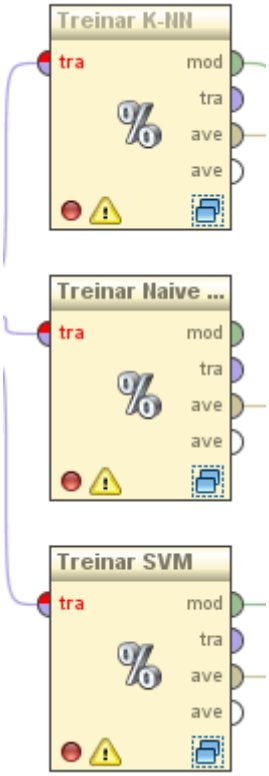
Fonte própria

O primeiro componente trata da transformação de todo o texto para minúsculas, facilitando assim futuras comparações. O "*Tokenize*" é o elemento responsável por separar o texto em *tokens*, estando configurado para assegurar essa quebra por caracteres que não sejam letras, ou seja, qualquer sinal de pontuação, espaço ou caracteres especiais serão identificados como separadores de palavras. Para uma primeira redução do número de termos pouco significativos, todos os *tokens* com menos de quatro ou com mais de vinte e cinco caracteres (palavras com estas dimensões não trazem estatisticamente relevância à análise (RapidMiner)) são excluídos do *bag-of-words*. O quarto componente introduz um parâmetro de ajuste, descrito anteriormente no ponto 4.3.1, que se baseia numa lista de expressões regulares para substituição de termos específicos a cada organização por termos mais genéricos. A fase seguinte,

depende do idioma, já que remove as *stop-words* do *bag-of-words*. Exemplo de uma lista de *stop-words* pode ser encontrado no anexo 2.

O próximo passo é a lematização (*stemming*), ou redução dos termos à sua forma base. Este passo é fortemente dependente do idioma. O último passo no processo de pré-processamento de texto é a geração de *n-grams*, ou seja, a geração de novos termos baseados na concatenação dos termos que surgem no texto a uma certa distância entre eles. Exemplo: os termos "sistema" e "operativo" podem fazer sentido numa análise enquanto um só termo, "sistema_operativo".

Este processo permitiu a obtenção de um *bag-of-words*. Este conjunto deverá agora ser processado para encontrar a matriz TF-IDF tal como descrito nos capítulos 2.3.2 e 2.3.5.4, e será com base nas frequências obtidas para cada termo que se farão as análises subsequentes. Neste pressuposto e no modelo definido, termos que ocorram num número pouco significativo de casos, ou em praticamente todos os casos são descartados por não terem um significado expressivo (frequências muito elevadas representam termos que não acrescentam valor à distinção entre as categorias).

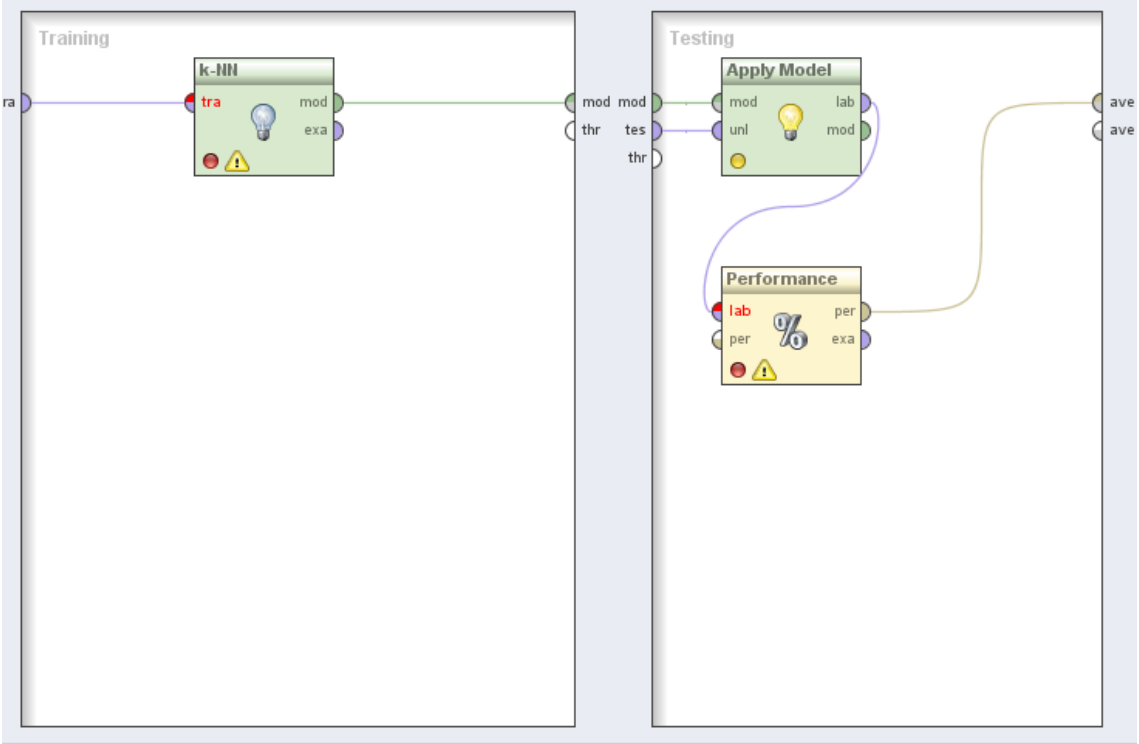


Encontrando-se o texto preparado, *tokenizado* e identificada a matriz TF-IDF, segue-se a fase de aprendizagem. Para tal, os casos (descrição de cada pedido ou incidente) são separados em dois conjuntos (de dimensão ¾ e ¼ do total), de forma aleatória: conjunto de aprendizagem e conjunto de teste, sendo aplicado o modelo de aprendizagem ao primeiro, para posteriormente ser aferido a sua eficácia no segundo. Na fase de aprendizagem o modelo comporta o teste com as principais metodologias:

- k-NN
- Naive Bayes
- SVM

Assim, são treinados os modelos para cada um dos algoritmos e obtidos resultados (ilustração 15). O algoritmo que obtiver melhor performance para os casos avaliados, será o elegido enquanto algoritmo para o sistema de inteligência semântica de triagem.

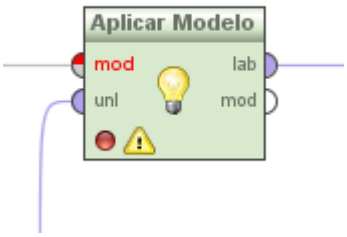
Ilustração 15 - Exemplo do treino do modelo com k-NN (RapidMiner)



Fonte própria

Consoante o algoritmo a utilizar, diferentes parâmetros podem ser ajustados por forma a otimizar a performance do modelo. Por exemplo, no caso do k-NN, o parâmetro k pode ser afinado.

Segue-se a fase de teste e análise de performance, onde o modelo é aplicado e obtidos os resultados, repetindo-se estas fases até se obter o valor máximo da performance do modelo.



Afinado o modelo que maximize a performance, o processo (*workflow* no Rapidminer) poderá ser finalizado e gravado para posterior utilização na componente de integração a desenvolver (interface Java), ou usado diretamente com recurso à exportação do

RapidMiner, ou aos utilitários disponibilizados por esta ferramenta para alteração direta em bases de dados. Este processo, uma vez finalizado, não carecerá de alterações a menos que surjam alterações significativas na organização.

O capítulo seguinte descreve a aplicação do modelo a um caso de estudo concreto.

5 Estudo de Caso

Encontrando-se o modelo definido, importa agora testar a sua robustez e aplicabilidade, usando-se para tal um caso real, a unidade de TI do Município de Oeiras.

5.1 Caraterização da Organização Objeto de Estudo de Caso

O Município de Oeiras dispõe de uma unidade de TI que se destaca dos demais municípios em termos de avanço tecnológico, primando por uma adoção visionária de novas tecnologias emergentes. Essa realidade de TI é aplicada sempre que possível na gestão do contacto com o munícipe, e colocada ao serviço do munícipe, tendo como alguns exemplos a disponibilização de locais com internet gratuita, como por exemplo as praias do concelho. Esta unidade de suporte de TI é responsável por assegurar o bom funcionamento destes recursos mais expostos, mas tem também como missão assegurar serviços críticos, como sejam os inerentes à segurança (Polícia Municipal e Proteção Civil nas suas redes de radiocomunicação e videovigilância) e postos de trabalho em geral. Por isto torna-se critico otimizar as estruturas de suporte, alocando o tempo recuperado na realização de outras tarefas importantes.

5.1.1 Análise das Tecnologias de Informação Existentes no Município

No caso em análise, a função de suporte permite assegurar uma infraestrutura equivalente a uma empresa de segmento médio. Em termos de dimensão, a estrutura a suportar é composta por dois centros de dados dispersos, uma rede de dados de banda larga a interligar mais de 60 locais com topologias de rede diferenciadas consoante as necessidades específicas de cada local, dos quais se incluem jardins-de-infância e escolas, um parque informático com mais de 1400 computadores, mais de 300 equipamentos de impressão e digitalização, mais de 1000 terminais de voz fixa IP, cerca de 450 telemóveis e mais de 140 quadros interativos.

5.1.2 Estatísticas de *Service Desk*

Após apurar os requisitos aplicacionais principais, tais como, garantir uma relação mais direta entre o canal de apoio técnico (*Service Desk*) e os utilizadores finais, gestão das intervenções por equipas através da categorização da ordem de serviço por parte do utilizador, a solução que se encontra implementada é o GLPI (*Gestion Libre de Parc*

Informatique), *software* que além de se encontrar disponível sob a forma de "código aberto" (*open source*) e consequentemente sem custos de utilização, cumpre com os requisitos elencados e que vão de acordo às normas e boas práticas instigadas pela ITIL. Sendo um *software* de *open source* facilita também o processo de integração com as tecnologias padronizadas pelo modelo.

A unidade de suporte de TI do município de Oeiras recebe uma média de 4.500 solicitações anuais, distribuídas de acordo com a tabela 11. Desta forma, constata-se que as categorias que obtêm mais solicitações são as de "*Software"* e *"Hardware"*, logo seguidas por "Rede" e "Impressão".

Tabela 11 - Distribuição das solicitações por categorias

Categoria	Distribuição
Software	28%
Impressão	11%
Hardware	26%
Rede	12%
Sistemas	9%
Correio Eletrónico	5%
Telefone	9%

Dados recolhidos no estudo de caso

Por observação do registo associado a cada solicitação na aplicação GLPI, constata-se que o tempo que decorre entre a receção do pedido e a atribuição à equipa respetiva (tempo de resposta) é em média de 1h15, ou seja, são consumidos anualmente em média cerca de 5625h[6] apenas na fase de triagem de pedidos, tempo esse em que os pedidos estão a aguardar por atribuição a uma equipa e o utilizador aguarda por uma resolução.

5.2 Aplicação do Modelo

Para a aplicação do modelo, foram seguidos os passos descritos no capítulo 4.3, iniciando-se pela recolha e preparação dos casos, separação dos casos obtidos em dois grupos (aprendizagem e teste), preparação e pré-processamento dos textos, treino do

[6] Valor obtido pela multiplicação do número de solicitações (4500) pelo tempo médio necessário à triagem de uma solicitação (1h15). Este cálculo representa um valor acumulado.

modelo (com ajuste dos parâmetros de otimização) e posterior validação do modelo treinado com os casos de teste. Os detalhes mais significantes desses passos serão dados a conhecer nos pontos seguintes.

5.2.1 Recolha e Preparação dos Dados

Tal como referido, os dados inerentes às solicitações de TI encontram-se armazenados na aplicação GLPI. Essa aplicação é desenvolvida em PHP e suportada pelo motor de base de dados *MySQL*. Para a obtenção dos dados, a pedido da organização, ficou estipulado que o processo deveria ser o menos intrusivo possível, pelo que se optou pela extração da informação necessária diretamente da base de dados para um ficheiro Excel, ficheiro esse que serviria de base aos restantes passos, evitando assim um acesso permanente à infraestrutura de servidores com as consequências inerentes a esta necessidade. Desta forma, foram exportados para um documento Excel diretamente da base de dados as solicitações (id, título, descrição e categoria) relativas ao ano de 2014, tendo-se obtido um total de 4.187 registos (casos).

Foi efetuada uma análise pelos registos (escolha aleatória) para tentar atestar da qualidade dos dados, tendo-se percebido que existem solicitações onde a descrição é muito escassa (curta: ex. "*PC-043241 não liga*") e praticamente pouco revela sobre o problema real, muito embora a taxa dessas ocorrências seja baixa.

Com base nessa observação, foi construída a lista de expressões regulares para substituição de termos, com vista à generalização de termos mais específicos à organização, tendo sido incluídos os seguintes:

Tabela 12 - Lista de substituições

Procurar (expressão regular)	**Substituir por**
pc-[0-9]{1,6}	computador
pc(\W)	computador$1
imp-[0-9]{1,6}	impressora
imp(\W)	impressora$1
printer(\W)	impressora$1
qi(\W)	quadro interativo$1
interactivo	interativo
mfc-[0-9]{1,6}	impressora
e-mail	email

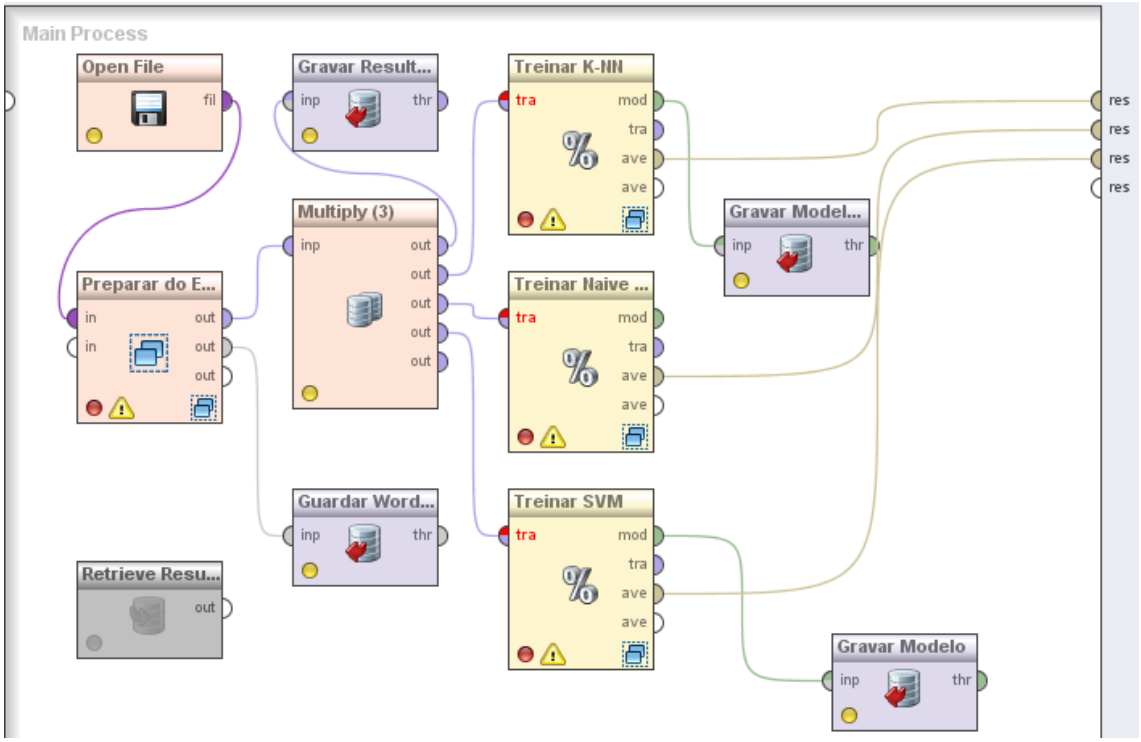
Fonte própria

Concluída a preparação dos documentos textuais, segue-se a fase de treino do modelo. Para tal, os documentos textuais foram selecionados de forma aleatória com a distribuição do total dos casos de ¾ para ¼, ficando ¾ (3095 casos) para a fase de treino do modelo e ¼ (1092 casos) para a fase de teste e validação.

5.2.2 Treino do Modelo

Para o treino do modelo foi programado no processo do RapidMiner a leitura de um ficheiro Excel com todos os casos de treino, e executado o processo (ilustração 16).

Ilustração 16 - Processo de treino do modelo (RapidMiner)



Fonte própria

Para a afinação do modelo, foram feitos testes com os diferentes algoritmos de treino, e ajustados os parâmetros de cada um por forma a maximizar a performance do modelo.

Para o teste usando o algoritmo k-NN foi-se variando o valor de k, tendo-se obtido os valores da tabela seguinte:

Tabela 13 - Treino do modelo (k-NN variação dos valores de k)

Valor de k	**1**	**2**	**3**	**5**	**10**	**15**	**20**	**30**	**50**	**100**
Performance	68,9%	68,9%	72,7%	77,3%	82,1%	**85,4%**	85,8%	86,0%	86,0%	85,2%
Melhoria face ao k anterior		0,0%	3,7%	4,6%	4,9%	3,3%	0,4%	0,2%	-0,1%	-0,7%

Fonte própria

Da análise dos valores da performance e respetiva melhoria com a variação de k, constata-se que para $k < 15$ os valores da performance aumentam bastante acima de 1% face ao valor anterior. Para $k > 15$, essa variação reduz para valores inferiores a

1%, chegando mesmo a valores de performance piores para $k > 50$. Assim, o valor ótimo encontrado para k foi de 15, obtendo-se uma performance global do modelo de 85,43% (± 0.55%) na fase de treino, como se pode observar na ilustração seguinte.

Ilustração 17 – Treino do modelo (Performance Vector k-NN - Rapid Miner)

accuracy: 85.43% +/- 0.55% (mikro: 85.43%)

	true Hardware	true Email	true Software	true Rede	true Telefone	true Sistemas	true Impressão	class precision
pred. Hardware	585	4	33	28	8	22	14	84.29%
pred. Email	8	310	21	7	1	12	2	85.87%
pred. Software	13	15	701	17	9	35	6	88.07%
pred. Rede	13	5	14	240	5	10	10	80.81%
pred. Telefone	8	2	5	6	132	3	1	84.08%
pred. Sistemas	5	18	29	18	6	291	11	76.98%
pred. Impressão	3	1	14	2	3	4	385	93.45%
class recall	92.13%	87.32%	85.80%	75.47%	80.49%	77.19%	89.74%	

Fonte própria – Extraído do RapidMiner

Verifica-se que apesar de o modelo obter uma performance satisfatória, existem ainda bastantes situações que se traduzem em categorizações erradas, sendo a classe "Sistemas" a que mais contribuir de forma negativa para o resultado final. O modelo necessitou de 1 minutos 45 segundos para obtenção dos resultados.

Foram assegurados os testes com o mesmo conjunto de dados de teste recorrendo ao algoritmo de SVM. Este algoritmo torna-se bastante mais lento, tendo necessitado de mais de 16 minutos e 26 segundos para apresentar os resultados na ilustração 18. Este modelo obteve assim uma performance de 85,75% (± 1,77%), resultados bastante semelhantes aos obtidos pelo k-NN (85,43%), tendo demorado cerca de 8 vezes mais tempo. Além desta desvantagem, os resultados apontam para problemas de precisão numa das classes (Telefone, com menos de 60% de precisão), pelo que a aplicação deste algoritmo, não é recomendável neste caso.

Ilustração 18 – Treino do modelo (Performance Vector SVM - Rapid Miner)

accuracy: 85.75% +/- 1.77% (mikro: 85.75%)

	true Hardware	true Email	true Software	true Rede	true Telefone	true Sistemas	true Impressão	class precision
pred. Hardware	571	4	21	19	4	10	8	89.64%
pred. Email	5	301	9	6	1	17	1	88.53%
pred. Software	18	15	740	19	9	43	6	87.06%
pred. Rede	9	10	9	246	9	12	9	80.92%
pred. Telefone	19	9	15	18	135	28	2	59.73%
pred. Sistemas	8	13	13	9	3	260	2	84.42%
pred. Impressão	5	3	10	1	3	7	401	93.26%
class recall	89.92%	84.79%	90.58%	77.36%	82.32%	68.97%	93.47%	

Fonte própria – Extraído do RapidMiner

Relativamente ao treino recorrendo ao método de *Naive-Bayes*, constata-se que os resultados obtidos (ilustração 19) são também piores que os obtidos por qualquer dos algoritmos testados anteriormente, com um valor de 75,61% (± 1,79%), muito embora o tempo de execução seja excecionalmente rápido, treinando o modelo em menos de 45 segundos.

Ilustração 19 – Treino do modelo (Performance Vector Naive-Bayes - RapidMiner)

accuracy: 75.61% +/- 1.79% (mikro: 75.61%)

	true Hardware	true Email	true Software	true Rede	true Telefone	true Sistemas	true Impressão	class precision
pred. Hardware	501	4	18	24	14	10	9	86.38%
pred. Email	8	265	43	14	10	41	5	68.65%
pred. Software	23	22	625	12	8	19	10	86.93%
pred. Rede	29	16	35	220	18	13	7	65.09%
pred. Telefone	42	12	16	28	103	26	6	44.21%
pred. Sistemas	15	32	52	13	6	248	14	65.26%
pred. Impressão	17	4	28	7	5	20	378	82.35%
class recall	78.90%	74.65%	76.50%	69.18%	62.80%	65.78%	88.11%	

Fonte própria – Extraído do RapidMiner

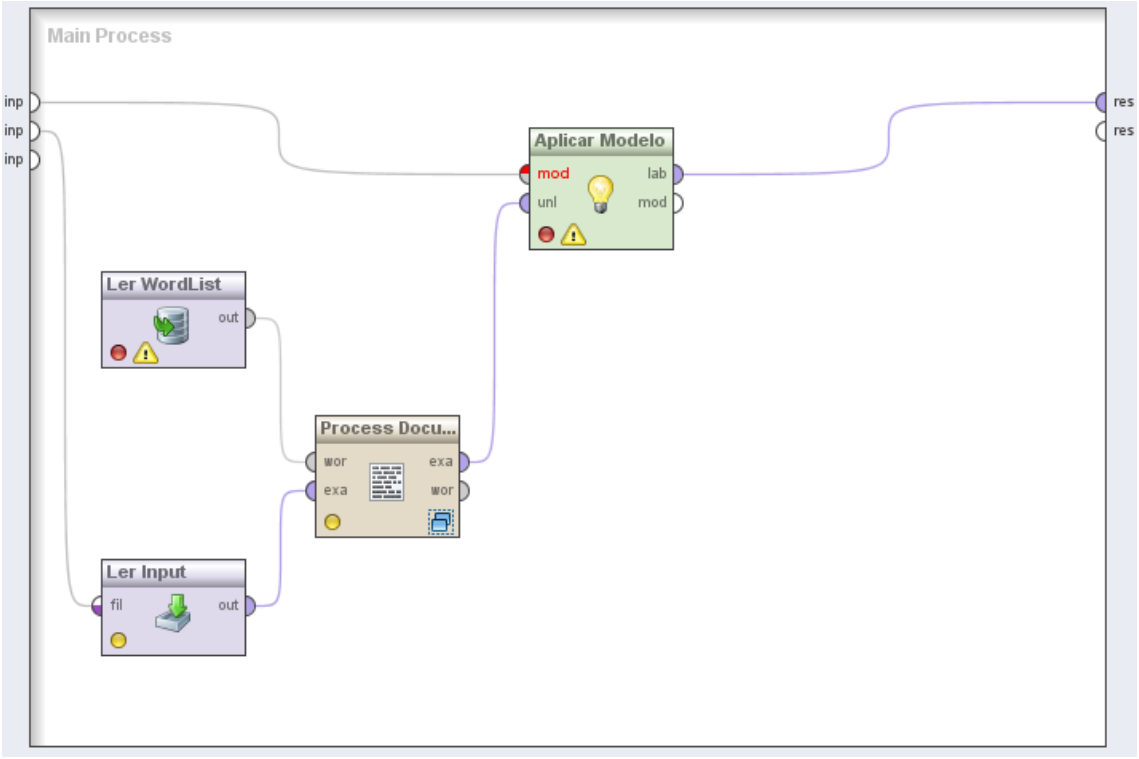
Assim, o algoritmo mais adequado, numa relação de precisão/tempo de execução, neste estudo de caso será o **k-NN** com um valor de $k = 15$.

5.2.3 Teste do Modelo

Para o teste do modelo foi concebido o processo do RapidMiner apresentado na ilustração 20. Para o teste o RapidMiner necessita de obter a lista de palavras (*bag-of-words*) gerada para pelo treino do modelo, bem como os casos de teste. O pré-processamento efetuado aos casos de treino é igualmente aplicado aos casos de teste,

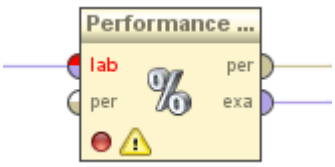
partindo posteriormente esse conjunto (já em formato de matriz TF-IDF marcada) para o teste do modelo.

Ilustração 20 – Processo de teste (RapidMiner)



Fonte própria – Extraído do RapidMiner

Após o teste do modelo, é calculada a matriz de performance recorrendo ao componente exibido à direita.



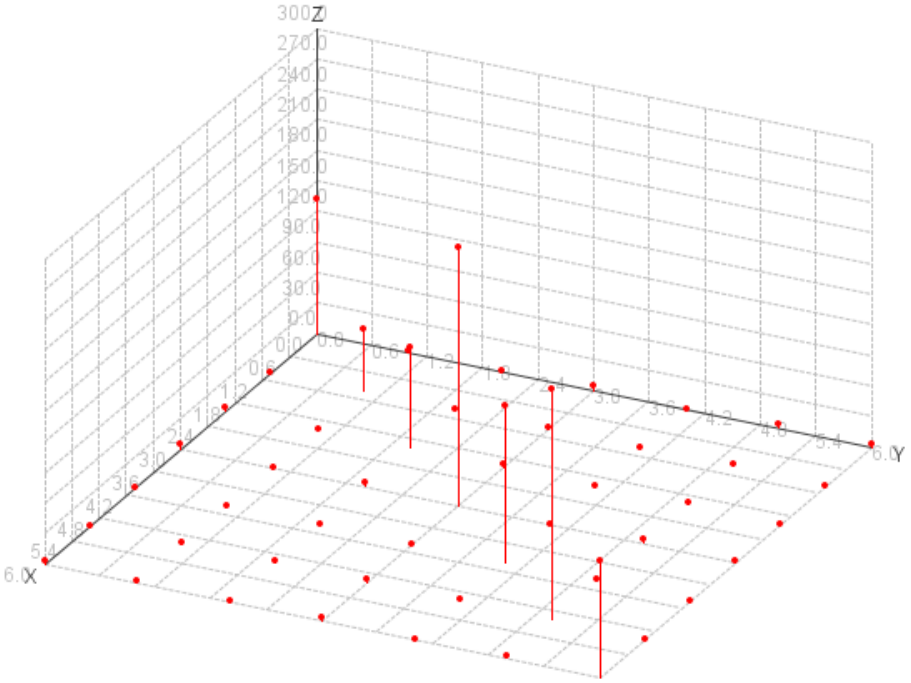
O resultado do cálculo da matriz de performance é o apresentado na ilustração 21, tendo sido obtido um valor de 95,15% de precisão do modelo, significando que apenas 4,85% dos casos serão categorizados erradamente ou não serão categorizados.

Ilustração 21 – Teste do modelo (Performance Vector - RapidMiner)

Table View | Plot View

accuracy: 95.15%

	true Sistemas	true Telefone	true Rede	true Software	true Impressão	true Hardware	true Email	class precision
pred. Sistemas	132	0	2	4	0	0	2	94.29%
pred. Telefone	1	60	0	0	0	2	1	93.75%
pred. Rede	0	1	99	4	1	2	1	91.67%
pred. Software	4	1	2	253	0	2	2	95.83%
pred. Impressão	0	0	0	0	154	2	0	98.72%
pred. Hardware	4	2	2	3	2	227	2	93.80%
pred. Email	3	0	0	1	0	0	114	96.61%
class recall	91.67%	93.75%	94.29%	95.47%	98.09%	96.60%	93.44%	



Fonte própria – Extraído do RapidMiner

No capítulo seguinte são analisados os resultados obtidos relativamente à aplicação do modelo.

5.3 Resultados Obtidos

Após o teste do modelo, foram analisados aleatoriamente os casos onde se constataram existir divergências entre a categoria apontada pelo modelo e a real categoria do caso, tendo-se chegado a algumas causas possíveis:

- **Descrição do pedido contempla múltiplas categorias**. O modelo caso tenha igual grau de certeza para categorias diferentes, atribui a primeira. Ex: "*Computador não liga e impressora não imprime*". Neste caso o modelo não deverá atribuir uma categoria, aguardando por categorização manual;
- **Triagens manuais incorretas**: foram encontrados casos (residuais) onde a triagem era incorreta, estando atribuídas categorias que não correspondiam à descrição textual. Talvez essas categorias tivessem sido alteradas após contacto telefónico, mas cuja alteração não foi refletida na aplicação GLPI.
- **As assinaturas utilizadas no final das mensagens de correio eletrónico** causam confusão ao modelo, já que é tradicional a colocação de extensões e números de telefone, o que confunde a categorização sendo frequente a atribuição da categoria "Telefones" nestes casos.

Ainda com estas situações, considera-se que o modelo é aplicável, e permitirá reduzir em 95% o esforço no processo de triagem e categorização de pedidos.

Atendendo ao valor calculado no capítulo 5.1.2 para as horas anuais despendidas no processo de categorização (5625h), podemos estimar que serão poupadas cerca de 5343h (já que 95% dos 4500 pedidos não serão categorizados manualmente) pela aplicação do modelo, e após a organização deter a confiança necessária no mesmo, o processo de atribuição às equipas poderá também deixar de ser validado.

Assim, considera-se que o modelo será uma mais-valia a aplicar traduzindo-se em vantagens a curto prazo, contribuindo não só para reforçar a confiança na equipa de suporte de TI e para o "Service Desk", mas também para melhorar a organização como um todo, ao reduzir os tempos de indisponibilidade de recursos de TI.

6 Conclusões e Trabalhos Futuros

6.1 Conclusões

A gestão do conhecimento, através da prospeção de informação, abarca um potencial de inovação tão extenso quanto o empenho e envolvimento que as organizações lhe disponibilizem. São reconhecidas as vantagens na criação de uma cultura organizacional, pelos motivos mais simples, desde a retenção do conhecimento dentro da organização quando os recursos humanos a abandonam, aos mais elaborados, de reaproveitamento do conhecimento para potenciar novas soluções e otimizar processos já existentes.

A crescente utilização de recursos suportados por tecnologias de informação e comunicação, funcionam, por um lado enquanto mecanismo facilitador da gestão do conhecimento, mas por outro são também o focos e origem de problemas e necessidade de suporte para assegurar para o correto funcionamento da organização.

É nesta gestão que o trabalho se centrou e onde se propõe a aplicar medidas suportadas na gestão do conhecimento para melhorar os processos de suporte existentes, reutilizando conhecimento organizacional já adquirido.

Para o efeito, e com o intuito de alcançar os objetivos específicos traçados, as tarefas inerentes à presente dissertação tiveram início com a tradicional recolha e prospeção de informação, tendo sido verificada a sua relevância e importância, bem como se a mesma era suficiente como base para a fundamentação necessária a esta redação. Constatando-se que alguns dos pontos careciam de maior sustentabilidade, nomeadamente o objetivo de "*Identificar e classificar o perfil dos serviços de TI relevantes*", existiu necessidade de conduzir um questionário focado nessa matéria, complementando assim com a informação secundária existente. Relativamente ao segundo objetivo, a análise do "... *estado da arte relativamente aos modelos de inteligência semântica aplicáveis*", e tendo por base a prospeção de informação nestas matérias, foram eleitos os métodos considerados mais adequados a esta pretensão, chegando-se desta forma ao terceiro objetivo, que se caracteriza pela criação de um modelo de inteligência semântica para

assistir uma implementação de ITIL. Este objetivo foi conseguido recorrendo às estratégias de "*Ground Theory*" e "*Secondary Data*", o que conduziu à criação de um modelo de inteligência semântica genérico e aplicável a uma implementação ITIL. Neste cenário, foi concebido um modelo de categorização e triagem automática de pedidos de solicitação que visa dotar a organização de uma ferramenta que permite melhorar significativamente o suporte de TI, tendo como requisito para a correta aplicação, a adoção de boas práticas da ITIL, que servirão enquanto pilar basilar do processo dando garantias do seu bom funcionamento.

O modelo foi concebido e testado, para posteriormente ser aplicado a um caso específico, completando assim o quarto e último objetivo da investigação, recorrendo à metodologia de investigação de "*Action Research*" e "*Grounded Theory*", onde, pelos resultados obtidos se pode constatar que a solução encontrada é proveitosa, muito embora careça de atualizações e desenvolvimentos, adaptando-a à dinâmica da organização onde se insere, considerando-se assim os objetivos cumpridos e encontrada a resposta à pergunta de investigação formulada.

6.2 Restrições ao Estudo

O desenvolvimento do modelo de inteligência artificial para uma implementação ITIL apresentou como principal restrição a ausência de casos reais com todas as variáveis e detalhe necessário que permitam uma recolha efetiva de informação de treino.

A pesquisa de casos e informação estatística relevante inerente à gestão de TI, incidindo especificamente sobre o "*Service Desk*", foi também um processo complexo, não existindo informação detalhada sobre esta realidade, motivo pelo qual foi necessário formular um questionário.

Relativamente à aplicação do questionário, e concretamente ao número de respostas obtidas, muito embora se considere suficiente, poderia ter sido superior. Esta mudança poder-se-ia traduzir em informação genericamente mais representativa da realidade, o que tornaria o estudo ainda mais adequado às organizações que dele venham a necessitar.

Teria sido importante também, no estudo do caso, ter avançado para métodos de implementação do modelo integrados com a aplicação de suporte existente, isto é, a aplicação ficar totalmente automatizada na componente de triagem. Esta automatização poderia ser assegurada com recurso a mecanismos de partilha de informação, como por exemplo, w*ebservices*.

O estudo necessita, no geral, de maior maturação, de mais validações e de maior esforço no sentido de limar algumas arestas, muito embora represente já uma base de solução para problemas complexos, sendo necessário dar continuidade ao trabalho desenvolvido.

6.3 Melhorias Futuras

O foco desta dissertação centrou-se na criação de um modelo de inteligência semântica para assistir um processo de ITIL. Apesar de os objetivos terem sido atingidos, a sua utilização prática carece ainda de mecanismos de integração do modelo com as diferentes aplicações utilizadas pelas diferentes organizações. Desta forma, uma melhoria futura a considerar, seria a criação de uma *framework* que permitisse a integração com aplicações existentes ou, em última análise, a construção de um portal mais simples que permitisse digitar ou colar texto (descritivo da solicitação) e que permitisse com base no mesmo identificar a categoria respetiva. Desta forma, este módulo poderia ser integrado quase transparentemente nas aplicações (sobretudo nas assentes em tecnologias web), podendo ficar a tarefa da categorização tratada por exemplo no momento da criação da solicitação.

Outro aspeto de possível melhoria prende-se com a possibilidade de realimentar o modelo a cada novo pedido ou incidente que é categorizado, garantindo assim um modelo em constante aprendizagem e evolução, adaptando-se a novos termos e tecnologias.

Referências Bibliográficas

Abramowicz, Witold et all. (2007). *A Need for Business Assessment of Semantic Web.* Australia: Springer.

Addy, R. (2007). ITIL – Holy Grail or Poisoned Chalice? *Effective IT Service Management*.

Aguinis, H. (1993). Action Research and Scientific Method: Presumed Discrepencies and Actual Similarities. *Journal of Applied Behavioral Science, 29*, 416-431.

Alexe, G., Alex, S., Hammer, P. L., & Kogan. (2002). *A. Comprehesive VS. Comprehensible Classifiers in Logical Analysis of Data.* New Jersey, USA: RUTCOR Research Report - State University of New Jersey.

Andersen, J. P., Prause, J., & Silver, R. C. (2011). *Step-by-Step Guide to Using Secondary Data for Psychological Research.* Social and Personality Psychology Compass.

Andrew, J., & Rob, A. (2014). *Profiling IT support*. Obtido em 01 de 2015, de https://www.umich.edu/cases/2014/profilingitsupport-andrew-j-rob-andy.html

Axelos. (2012). *Glossário e abreviações ITIL - Português do Brasil.* AXELOS.

Axelos. (2015). *Axelos Website*. Obtido em Novembro de 2014 a Março de 2015 de 2014/2015, de https://www.axelos.com

Ayehu. (2015). 5 Reasons you should automate - Security Incident Responses.

Badham, R. J., & Sense, A. J. (2006). Spiralling Up or Spinning Out: A Guide for Reflecting on Action Research Practice. *International Journal of Social Research Methodology, 9*, 367-377.

Baeza-Yates, R., & Neto, B. R. (1999). *Modern Information Retrieval.* ACM Press.

Baskerville, R. L. (1999). Investigating Information Systems with Action Research. *Communications of the Association for Information Systems, 2*(3es), 4.

Bastos, V. M. (2006). *Ambiente de Descoberta de Conhecimento para a Língua Portuguesa.* Rio de Janeiro - Brasil: COPPE/UFRJ.

Benbasat, I., Goldstein, D., & Mead, M. (1987). The Case Research Strategy In Studies Of Information Systems.

Boog, B. W. (2003). The emancipatory character of action research, its history and the present state of the art. *Journal of Community and Applied Social Psychology, 13*, 426-438.

Boog, B., & al., e. (1996). *Theory of Action Research: With Special Reference to the Netherlands Tilburg.* The Netherlands: Tilbury University Press.

Boser, S. (2006). Ethics and power in community-campus partnerships for research. *Action Research, 4*, 9-21.

Cartlidge, A., Hanna, A., Rudd, C., Macfarlane, I., Windebank, J., & Rance, S. (2007). *An Introductory Overview of ITIL® V3 - The IT Infrastructure Library - Version 1.* The UK Chapter of the itSMF.

Chen¸Hsinchun, & Chiang, R. &. (2012). *Business intelligence and analytics: From big data to big impact.* New York: MIS.

Chisholm, R., & Elden, M. (1993). Features of Action Research. *Human Relations, 46.2*, 275-298.

Coelho, J. M. (2009). *Maturidade da Gestão de Serviço de TI – Parte II. Governação de TI.* itSMF.

Committee on Institutional Cooperation. (1997). Final Report - Incident Cost Analysis and Modeling Project. Committee on Institutional Cooperation.

Coy, M. (2006). This Morning I'm A Researcher, This Afternoon I'm An Outreach Worker: Ethical Dilemmas in Practitioner Research. *International Journal of Social Research Methodology, 9,* 419–431.

Creswell, J. W. (2003). *Research Design, Qualitative, Quantitative and Mixed Methods.* Sage Publications.

Crown. (2011). *ITIL® glossary and abbreviations - English.* Crown Copyright.

da Silva, J. F. (2003). *Extracção de Unidades Textuais, Agrupamento Caracterização e Classificação de Documentos.* Lisboa: Universidade Nova de Lisboa - Faculdade de Ciência e Tecnologia.

Das Graças Volpe Nunes, M., & Specia, L. (2004). *Desambiguação Lexical Automática de Sentido: Um Panorama.* Brasil: NILC - Núcleo Interistitucional de Linguística Computacional.

Davenport, T. H. (2000). *Working Knowledgeg.* Harvard Business of School.

Davenport, T. H., & Prusak, L. (1998). *Working Knowledge.* HBS Press-USA.

David, M. (2002). Problems of participation: the limits of action research. *International Journal of Social Research Methodology, 5*, 11-17.

Deming, W. (1986). *Out of the Crisis. MIT Center for Advanced Engineering Study.*

Dick, B. (2006). Action research literature. *Action Research, 4,* 439-458.

Drucker, P. (1999). *Desafios gerenciais para o século XXI.* São Paulo: Pioneira.

Drucker, P. (2006). *Classic Drucker.* USA: Harvard Business Review.

Eden, C., & Huxham, C. (1996). Action research for management research. British Journal of Management.

Elden, M., & Chisholm, R. (1993). Emerging Varieties of Action Research: Introduction to the Special Issue. *Human Relations, 46.2*, 121-42.

Fagundes, E. M. (02 de 2010). *SlideShare - Gestão de Serviços de TIC.* Obtido em 04 de 2015, de http://pt.slideshare.net/emfagundes/gesto-de-servios-de-tic

Ferreira, A. P. (2011). *Implementação de processos da fase de operação de serviço do ITIL® em ambiente universitário: o caso do ISCTE-IUL.* ISCTE-IUL.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures And Algorithms.* Prentice Hall PTR.

Freitas, C. M., Chubachi, O. M., Luzzardi, P. R., & Cava, R. A. (2001). *Introdução à Visualização de Informações.* Brasil: RITA - UFRGS.

Gershon, N. (1977). *Information Visualization.* USA: IEEE - Computer Graphics and Application.

Gillham, B. (2001). *Case Study Research Methods.* London, New York: Continuum.

Hagel, J., & J.S, B. (2004). *TI Flexível, a melhor estratégia.* HSM Management.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques.* Academic Press.

Haykin, S. (1999). *Neural Networks: a comprehensive foundations.* Prentice Hall, Inc.

Haykin, S. (1999). *Neural Networks: a comprehensive foundations.* Prentice Hall, Inc.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum.*

ITIL Service Management. (11 de 2013). Obtido em 02 de 2015, de http://itservicemngmt.blogspot.pt

itSMF. (2007). *An Introductory Overview of ITIL® V3. A high-level overview of the IT INFRASTRUCTURE LIBRARY.* © Crown.

Jacquemin, C. (1996). *A Symbolic and Surgical Acquisition of Terms Through Variation.* Paris, França: Institut de Recherches en Informatique de Nantes (IRIN).

Jacquemin, C. (1996). *A Symbolic and Surgical Acquisition of Terms Through Variation.* Paris: Institut de Recherches en Informatique de Nantes.

Janowicz, K., Li, N., Bodenreider, O., & Kiss, E. (2012). The Use of Semantic Web Technologies for Decision Support - A Survey. *IOS Press*.

Korth, H. F., & Silbertchatz, A. (1993). *Sistema de Gerenciamento de Banco de Dados.* Makron Books.

Liu, B. (2011). *Web Data Mining - Exploring Hyperlinks, Contents and Usage Data - Second Edition.* Chicago: Springer.

Loh, S., Amaral, L. A., Wives, L. K., & de Oliveira, J. P. (2006). *Descoberta de Conhecimento em Textos através da Análise de Sequências Temporais.* Florianópolis, Santa Catarina, Brasil: WAAMD - XXI Simpósio Brasileiro de Banco de Dados.

Marrone, M., & Kolbe , L. M. (2011). Impact of IT Service Management Frameworks on the IT Organization. *Business & Information Systems Engineering, 3*(1), 5-18.

Martins, J. C., & Belfo, F. (2009). Métodos de Investigação Qualitativa - Estudos de Casos na Investigação em Sistemas de Informação.

Mazza, R. (2004). *Introduction to Information Visualisation.* USA: Faculty of Communication Sciences.

Merriam, S. (1998). *Qualitative research and case study applications in education.* San Francisco: Jossey-Bass.

Merriam-Webster Dictionary - Semiotics. (s.d.). (An Encyclopædia Britannica Company) Obtido em 06 de 2015, de http://www.merriam-webster.com/dictionary/semiotics

Miles, M., & Huberman, M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook (2nd ed.).* Newbury Park: Sage Publications.

Nelson, K., & Aaron, S. (2008). *Change Management No Longer Optional for Clients.* Management Consulting News.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company.* Oxford University Press, Inc.

Norris, D. (11 de 2013). *RapidMiner - a potential game changer*. Obtido em 07 de 2015, de http://www.bloorresearch.com/analysis/rapidminer-a-potential-game-changer/

Pardo, T. A., & das Graças Volpe Nunes. (2003). *M. Análise de Discurso: Teorias Discursivas e Aplicações em Processamento de Líguas Naturais.* São Paulo: NILC-ICMP-USP, São Carlos, São Paulo, Brasil.

Pardo, T. A., & das Graças Volpe Nunes, M. (2003). *Análise de Discurso: Teorias Discursivas e Aplicações em Processamento de Líguas Naturais.* São Carlos, São Paulo, Brasil: NILC-ICMP-USP.

Perks, C., & Beveridge, T. (2003). *Guide to Enterprise IT Architecture.* Springer.

Pink Elephant. (2008). The Benefits of ITIL®.

RapidMiner. (s.d.). *RapidMiner Forum*. (RapidMiner) Obtido em 2015, de http://forum.rapid-i.com/

Reason, P., & H, B. (2008). *Handbook of Action Research: Participative inquiry and practice 2nd edition.* London: Sage Publications.

Rezende, S. O. (2003). *Sistemas Inteligentes - Fundamentos e Aplicações (RECOPE-IA - Rede Cooperativa de Pesquisa em Inteligência Artificial).* Brasil: Editora Manole.

Roche, C. (2003). *Ontology: A survey.* France: University of Savoie - IFAC.

Ross, K. (02 de 2015). *2014 MEGA BREACHES: 5 KEY TAKEAWAYS.* Obtido em 04 de 2015, de http://blog.blackbox.com/technology/tag/network-security/

Russel, P., & Stuart, N. (2004). *Artificial Intelligence. A Modern Approach.* Prentice-Hall.

Salton, G., & A., W. (1975). *Information Retrieval and Language Processing - A Vector Space Model for Automatic Indexing.* C.A. Montgomery.

Sannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technincal Journal*, 379–423,623–656.

Sardinha, T. B. (2004). *Linguística de Corpus.* São Paulo: Editora Manole.

Saunders, M., Lewis, P., & Thornhill, A. (2003). *Research Methods for Business Students.* Pearson.

Seale, C. (2000). *Introduction to qualitative methods.* Thousand Oaks, CA: Sage.

Setzer, V. W. (Dez de 1999). Dado, Informação, Conhecimento e Competência. *DataGramaZero - Revista de Ciência da Informação - n. zero*. Obtido em 05 de 2015, de http://www.dgz.org.br/dez99/Art_01.htm

Silva, J. M., Carvalho, C. L., & Ambrósio, A. P. (2005). *Uma Arquitetura para Desenvolvimento da Web Semântica Baseada em Comunidades Virtuais de Prática.* Goiás: Universidade Federal de Goiás, Instituto de Informática, Brasil.

Stake, R. (1995). *The Art of Case Study Research.* Thousand Oaks, CA: Sage Publications.

Tan, A.-H. (1999). *Text Mining: The state of the art and the challenges.* Singapore: Kent Ridge Digital Labs.

Tan, C. C. (2006). The Theory and Practice of Change Management. Asian Business & Management.

Toussi, F. (09 de 2014). *HSQLDB Database Manager - 1.8.0.* Obtido em 02 de 2015, de http://hsqldb.org/doc/2.0/util-guide/dbm-chapt.html

VirginiaTech. (2012-2013). *Network Infrastructure & Services - Annual Report.* Blacksburg: VirginiaTech. Obtido em 03 de 2015

W., H., & van Bon, J. (2008). *Functions and process in IT management. The Process Management Matrix.*

Ward, J., & Peppard, J. (2002). *Strategic Planning for Information Systems* (3rd Edition ed.). Wiley.

Wives, L. K. (2004). *Utilizando Conceitos como Descritores de Textos para o Processo de Identificação de Conglomerados (Clustering) de Documentos.* Rio Grande do Sul - Brasil: Universidade Federal do Rio Grande do Sul.

Wives, L. K., & Loh. (2006). *S. Tecnologias de Descoberta de Conhecimento em Informações Textuais (Ênfase em Agrupamento de Informações).* Rio Grande do Sul, Brasil: Universidade Federal do Rio Grande do Sul.

Yin, R. (1994). *Case Study Research: Design and Methods (Second ed.).* Thousand Oaks, CA: SAGE Publications.

Zisblat, J. (2008). O Impacto das Práticas ITIL na Flexibilidade Organizacional - Evidências Empíricas em uma Empresa Multinacional de TI. Rio de Janeiro: Fundação Getulio Vargas.

7 ANEXOS

Anexo 1 - Questionário sobre "Áreas de suporte nas TIC"

Areas de suporte nas Tecnologias de Informação e Comunicação (TIC)

O objectivo deste questionário sobre a gestão nas Tecnologias de Informação e Comunicação (TIC) é recolher informação relativa à importância das áreas de suporte relacionadas.

Este questionário foi elaborado com o intuito de recolher dados primários de suporte à dissertação "Modelo de Inteligência Semântica para uma Implementação ITIL", no âmbito do Mestrado em Gestão de Tecnologias e Sistemas de Informação.

As suas respostas serão tratadas de forma anónima.

* Required

Qual é a natureza da instituição onde trabalha? *

- Empresa
- Instituição sem fins lucrativos
- Administração pública

Qual a posição que ocupa na organização? *

- Utilizador: uso as TIC para execução do meu trabalho diário
- Operacional: o meu trabalho é garantir o correto funcionamento das TIC
- Gestor: participo na elaboração das políticas de TIC
- Decisor: tenho poder de decisão na definição e implementação das políticas de TIC

Das seguintes áreas de atuação TIC, atendendo à importância que cada representa na sua organização, escolha as 5 mais importantes:

1 - mais importante, 5 menos importante

	1	2	3	4	5
Administração de Redes	○	○	○	○	○
Sistemas de Impressão	○	○	○	○	○
Videovigilância	○	○	○	○	○
Incidentes e pedidos relacionados com Software	○	○	○	○	○
Telefone e Comunicações de Voz	○	○	○	○	○
Sistemas de Correio Eletrónico	○	○	○	○	○
Internet	○	○	○	○	○
Incidentes e pedidos relacionados com Hardware	○	○	○	○	○
Desenvolvimento de Software	○	○	○	○	○
Empréstimos de equipamento	○	○	○	○	○
Administração de Sistemas	○	○	○	○	○
Licenciamento de Software	○	○	○	○	○

Qual das seguintes medidas considera de implementação mais prioritária com vista à melhoria da qualidade no apoio ao utilizador: *

- Melhorar a eficiência na resolução de incidentes e pedidos
- Melhorar os softwares usados no apoio ao utilizador
- Melhorar os equipamentos usados para o apoio ao utilizador 1

Anexo 2 - Stop-words (Português)

de	muito	elas	aquela	houve
a	nos	qual	aqueles	houvemos
o	já	nós	aquelas	houveram
que	eu	lhe	isto	houvera
e	também	deles	aquilo	houvéramos
do	só	essas	estou	haja
da	pelo	esses	está	hajamos
em	pela	pelas	estamos	hajam
um	até	este	estão	houvesse
para	isso	dele	estive	houvéssemos
é	ela	tu	esteve	houvessem
com	entre	te	estivemos	houver
não	depois	vocês	estiveram	houvermos
uma	sem	vos	estava	houverem
os	mesmo	lhes	estávamos	houverei
no	aos	meus	estavam	houvera
se	seus	minhas	estivera	houveremos
na	quem	teu	estivéramos	houverão
por	nas	tua	esteja	houveria
mais	me	teus	estejamos	houveríamos
as	esse	tuas	estejam	houveriam
dos	eles	nosso	estivesse	sou
como	você	nossa	estivéssemos	somos
mas	essa	nossos	estivessem	são
ao	num	nossas	estiver	era
ele	nem		estivermos	éramos
das	suas	dela	estiverem	eram
à	meu	delas	hei	fui
seu	às	esta	há	foi
sua	minha	estes	havemos	fomos
ou	numa	estas	hão	
quando	pelos	aquele		

foram	formos	tem	tivera	tiverem
fora	forem	temos	tivéramos	terei
fôramos	serei	tém	tenha	terá
seja	será	tinha	tenhamos	teremos
sejamos	seremos	tínhamos	tenham	terão
sejam	serão	tinham	tivesse	teria
fosse	seria	tive	tivéssemos	teríamos
fôssemos	seríamos	teve	tivessem	teriam
fossem	seriam	tivemos	tiver	
for	tenho	tiveram	tivermos	

Anexo 3 – Excerto da matriz TF-IDF resultante do estudo de caso

ExampleSet (3095 examples, 2 special attributes, 1140 regular attributes) View Filter (3095 / 3095): all

Row No.	id	categoria	abaix_indic	abertur	abertur_fich...	abre	abri	abrir	abrir_ficheir	abro	aced	aced_aplic	aced_inte
	2014000001	Hardware	0	0	0	0	0	0	0	0	0	0	0
2	2014000002	Email	0	0	0	0	0	0	0	0	0	0	0
3	2014000004	Email	0	0	0	0	0	0	0	0	0	0	0
4	2014000005	Software	0	0	0	0	0	0	0	0	0.106	0	0
5	2014000006	Rede	0	0	0	0	0	0	0	0	0	0	0
6	2014000009	Telefone	0	0	0	0	0	0	0	0	0	0	0
7	2014000010	Sistemas	0	0	0	0	0	0	0	0	0	0	0
8	2014000011	Rede	0	0	0	0	0	0	0	0	0	0	0
9	2014000014	Email	0	0	0	0	0	0	0	0	0	0	0
10	2014000015	Software	0	0	0	0	0	0	0	0	0.101	0.181	0
11	2014000016	Software	0	0	0	0	0	0	0	0	0	0	0
12	2014000019	Software	0	0	0	0	0	0	0	0	0	0	0
13	2014000020	Email	0	0	0	0	0	0	0	0	0	0	0
14	2014000021	Hardware	0	0	0	0	0	0	0	0	0	0	0
15	2014000024	Hardware	0	0	0	0	0	0	0	0	0	0	0
16	2014000027	Software	0	0	0	0	0	0	0	0	0	0	0
17	2014000028	Software	0	0	0	0	0	0	0	0	0	0	0
18	2014000030	Software	0	0	0	0	0	0	0	0	0	0	0