

Faculdade de Economia da Universidade do Porto
Mestrado em Análise de dados e Sistemas de apoio à decisão

Extracção da informação de tabelas contidas em texto
- uma aplicação a Relatórios e Contas de empresas portuguesas -

Ana Cristina da Costa e Silva.

Orientadores: Professor Doutor Alípio Jorge e Professor Doutor Luís Torgo

Dissertação com vista à obtenção do grau de Mestre em Análise de dados e Sistemas de
apoio à decisão pela Faculdade de Economia da Universidade do Porto

Porto, 2002

Nota biográfica

Nascida em Angola, filha de Pai Português e Mãe Brasileira, Ana Cristina da Costa e Silva é uma entre muitos obrigados a deixar a sua terra natal, habitando no Porto desde os três anos. Enquanto estudante da Licenciatura em Gestão, participa no programa Erasmus, frequentando um semestre escolar na Handelshojsskole i Kolding – Dinamarca, e faz o seu estágio curricular nos escritórios de uma empresa de auditoria na Cidade do México, aonde vem a conhecer o seu marido. Regressada à terra Lusa, termina a Licenciatura em 1998 e começa a sua carreira profissional no Departamento de Estatística do Banco de Portugal, aonde o gosto pela descoberta de conhecimento escondido atrás dos dados mais veio a florescer.

Foi este gosto que a levou a optar por este Mestrado.

Agradecimentos

Gostaria de agradecer em primeiro lugar aos meus pais, Arlindo e Thereza, e marido, Angel, por me apoiarem e inspirarem ao longo da minha vida académica.

Quero também agradecer aos meus Professores, pelo esforço de transmissão dos seus conhecimentos, principalmente ao corpo docente deste Mestrado e muito especialmente aos meus orientadores, Professor Doutor Alípio Jorge e Professor Doutor Luís Torgo.

Mais agradeço aos meus chefes, Dr. Orlando Calião, Dr. António Garcia, Dr. Guilherme Pinto e Dra. Cecília Flores, pelo apoio e compreensão dados a este esforço.

Resumo

Actualmente, milhares de empresas na União Europeia são obrigadas a publicar, pelo menos anualmente, relatórios e contas descritivos da evolução económico-financeira da sua actividade. Esta informação é útil à tomada de decisões de diversos agentes económicos, sendo custosa a análise de números elevados de relatórios, pois a sua alimentação para bases de dados tem de ser feita manualmente. O objectivo desta tese é produzir um método que permita extrair, automaticamente, um conjunto de dados que, legalmente, tais documentos devem conter e que são apresentados em tabelas.

O primeiro passo que propomos à consecução deste objectivo consiste em classificar cada linha do relatório como “tem características de tabela” ou “não tem características de tabela”, levando em conta elementos que caracterizam a própria linha, a sua vizinhança, a página em que está inserida e o documento de que provém. Esta análise multi-nível permite ao método ser sensível aos estilos utilizados pelos autores para dispor a informação no papel, que é muito diversificado neste contexto marcado pela influência do marketing. Os resultados mostram que é possível seleccionar um conjunto reduzido de linhas que tem elevada probabilidade de conter a informação desejada.

Os restantes passos incluem: 1. localizar com precisão as tabelas nos documentos; 2. segmentar as respectivas células; 3. distinguir as células que contêm dados das que os caracterizam; 4. identificar as células que devem ser lidas conjuntamente; 5. interpretar os resultados, eventualmente recorrendo a conhecimento específico do contexto. As estratégias de resolução destas tarefas foram delineadas, sendo o conhecimento extraído em cada fase utilizado para refinar as decisões anteriores e basear as seguintes. Finalmente, as restrições impostas pelo contexto contabilístico podem ser utilizadas para testar a qualidade dos resultados e corrigir os erros que advenham de decisões incorrectas nas fases anteriores do processo, garantindo assim máxima fiabilidade à informação extraída.

Índice

1. INTRODUÇÃO

<u>1.1. Porquê extrair informação de relatórios e contas?</u>	<u>1</u>
<u>1.2. O reporte financeiro em Portugal</u>	<u>2</u>
<u>1.3. O reporte financeiro no futuro</u>	<u>3</u>

2. EXTRACÇÃO DE INFORMAÇÃO DE TABELAS CONTIDAS EM TEXTO

<u>2.1. O que é extracção de informação?</u>	<u>6</u>
<u>2.2. O estudo de tabelas</u>	<u>7</u>
2.2.1. Áreas de investigação que se interessam por tabelas	7
2.2.2. Modelos de representação de tabelas e tarefas associadas	8
<u>2.3. Tarefas na análise de tabelas</u>	<u>11</u>
2.3.1. Localização	11
2.3.1.1. Em ficheiros de imagem	11
2.3.1.2. Em texto simples	13
2.3.1.3. Em ficheiros etiquetados	15
2.3.2. Segmentação	17
2.3.3. Análise funcional	23
2.3.4. Análise estrutural	28
2.3.5. Interpretação	30
2.3.5.1. Obtenção das categorias em análise	30
2.3.5.2. Caracterização das relações entre células	31
2.3.5.3. Interpretação em contexto específico	32

3. UM MÉTODO DE LOCALIZAÇÃO DE TABELAS EM TEXTO

<u>3.1. A selecção da amostra de relatórios e contas a tratar</u>	<u>36</u>
<u>3.2. Breve caracterização do arranjo do espaço em relatórios e contas</u>	<u>38</u>
<u>3.3. Características desejáveis num sistema que trate relatórios e contas</u>	<u>42</u>

3.4. A primeira tarefa – localizar tabelas	43
3.4.1. Que variáveis independentes utilizar? _____	44
3.4.1.1. Descritivas da própria linha _____	44
3.4.1.2. Descritivas do documento de origem _____	45
3.4.1.3. Descritivas das linhas vizinhas _____	49
3.4.2. Construção do classificador _____	49
3.4.2.1. Variantes do método de classificação _____	58
3.4.3. Metodologia de avaliação dos resultados _____	61
3.4.3.1. No contexto específico – captar as tabelas interessantes _____	61
3.4.3.2. No contexto genérico – captar qualquer tabela _____	64
3.4.3.3. Conjunto de teste _____	65
3.4.4. Resultados experimentais no conjunto de treino _____	66
3.4.4.1. Avaliação em contexto específico _____	66
3.4.4.2. Avaliação em contexto genérico _____	74
3.4.5. Resultados experimentais no conjunto de teste _____	76
3.4.6. Trabalho futuro _____	77
3.4.7. Conclusão _____	78

4. PLANEAMENTO DO FUNCIONAMENTO DAS RESTANTES TAREFAS

4.1. Definição do modelo físico das tabelas, parte I - segmentação em colunas	81
4.1.1. Segmentação das linhas em células _____	81
4.1.2. Repartição das células pelas suas colunas _____	81
4.1.3. Validação do número total de colunas por tabela _____	81
4.2. Complementos à localização - eliminação das falsas tabelas que restam	81
4.3. Análise funcional	81
4.4. Definição do modelo físico das tabelas, parte II – células abarcadoras	81
4.4.1. Identificar células que abarcam mais que uma linha _____	81
4.4.2. Identificar células que abarcam mais que uma coluna _____	81
4.5. A construção da <i>template</i>	81
4.5.1. Oportunidades para a extracção em contexto contabilístico _____	81
4.5.2. Ameaças à extracção em contexto contabilístico _____	81
4.5.3. A construção da <i>template</i> _____	81

4.6. Ordenação das tabelas por grau de interesse	82
4.6.1. Um método apto em qualquer contexto	82
4.6.2. Aproveitando as oportunidades do contexto específico	82
4.6.2.1. Hipótese A	82
4.6.2.2. Hipótese B	82
4.7. Análise estrutural	82
4.8. Interpretação	82
4.8.1. Captação das relações partitivas	82
4.8.2. Identificação da unidade de medida	82
4.8.3. Verificação das relações entre tabelas e entre colunas da mesma tabela	82
 5. CONCLUSÃO	
 Apêndice 1 – teste de hipóteses	88
1.1. O número de espaços interiores é independente da pertença a uma tabela?	88
1.2. Relatórios diferentes apresentam diferentes distribuições de espaços internos?	90
1.3. As variáveis Tabela e Relatório são independentes?	92
Apêndice 2 – segmentação de linhas em células	94
 Referências bibliográficas	98
 Anexo 1 – distribuição de Espaços interiores	104
Anexo 2 – distribuição de Espaços interiores por documento	105
Anexo 3 – avaliação em contexto específico	111
Anexo 4 – avaliação em contexto genérico	113

Índice de figuras

Figura 1. Um exemplo de problema de extracção de informação.....	6
Figura 2. Relações genéricas de <i>input-output</i> entre as tarefas e modelos de representação associados às tabelas.....	10
Figura 3. Estruturas que compõem o modelo físico da tabela.....	18
Figura 4. Exemplo para ilustração do método de Hurst (2001).....	22
Figura 5. Detalhe da Sonae 1999 – texto anormalmente distribuído por colunas.	39
Figura 6. Detalhe da Brisa 2000 – gráficos, tabelas e texto inconsistentemente distribuídos na página.	39
Figura 7. Detalhe da EDP 1999 – várias linhas não preenchidas e Sonae 2000 – tabelas com poucas colunas.	39
Figura 8. Detalhes da Portucel 2000 - células de referência abarcando várias linhas <i>versus</i> hierarquia de conceitos.....	40
Figura 9. Passagem para texto do conteúdo da Figura 5. Detalhe da Sonae 1999 – texto anormalmente distribuído por colunas. (a).	41
Figura 10. Detalhe da Shell 1999 – tipo de letra transposta com erro para ASCII.....	41
Figura 11. A “divisória ideal” <i>versus</i> o número de linhas (a), a moda excluindo o zero (b) e o peso de zero espaços (c).....	48
Figura 12. Peso das linhas com tabelas em cada página, por tipo de tabela que a página contém.....	50
Figura 13. Peso dos candidatos a tabelas em cada página, por tipo de tabelas que a página contém.....	51
Figura 14. Escolha do método - fronteira de eficiência.....	66
Figura 15. Desempenho do método escolhido por relatório.....	67
Figura 16. Extracto de tabelas com Empresas incluídas na consolidação da EDP 2000.	71

Figura 17. Tipologia de erros do método base.....	75
Figura 18. Esquema das relações de <i>input-output</i> entre as diferentes fases de resolução do problema.....	83
Figura 19. Distribuição de espaços interiores no total da amostra.....	104
Figura 20. Proporção de tipos quando não há espaços interiores	104
Figura 21. Distribuição de espaços brancos interiores na Autoindustrial 1998.....	105
Figura 22. Distribuição de espaços brancos interiores na Autoindustrial 1999.....	105
Figura 23. Distribuição de espaços brancos interiores na Autoindustrial 2000.....	105
Figura 24. Distribuição de espaços brancos interiores na Brisa 1999.....	106
Figura 25. Distribuição de espaços brancos interiores na Brisa 2000.....	106
Figura 26. Distribuição de espaços brancos interiores na Cimpor 2000	106
Figura 27. Distribuição de espaços brancos interiores na EDP 1997.....	106
Figura 28. Distribuição de espaços brancos interiores na EDP 1998.....	107
Figura 29. Distribuição de espaços brancos interiores na EDP 1999.....	107
Figura 30. Distribuição de espaços brancos interiores na EDP 2000.....	107
Figura 31. Distribuição de espaços brancos interiores na Galpenergia 2000	107
Figura 32. Distribuição de espaços brancos interiores na Lusomundo 1999	108
Figura 33. Distribuição de espaços brancos interiores na Novabase 2000.....	108
Figura 34. Distribuição de espaços brancos interiores na Portucel 2000	108
Figura 35. Distribuição de espaços brancos interiores na RAR 2000.....	108
Figura 36. Distribuição de espaços brancos interiores na Shell 1999.....	109
Figura 37. Distribuição de espaços brancos interiores na Siemens 1999.....	109
Figura 38. Distribuição de espaços brancos interiores na Sonae 1999	109
Figura 39. Distribuição de espaços brancos interiores na Sonae 2000	109

Índice de tabelas

Tabela 1. Sumário das estratégias de localização de tabelas.	17
Tabela 2. Exemplo extraído de Green e Krishnamoorthy (1995a) para ilustração do seu método.....	19
Tabela 3. Exemplo para ilustração de Tupaj et al. (1996).....	19
Tabela 4. Exemplo para ilustração de Kieninger (1998).....	20
Tabela 5. Sumário das estratégias de reconhecimento de tabelas.....	23
Tabela 6. Exemplo extraído de Chen et al. (2000) para ilustração do funcionamento do seu método.....	27
Tabela 7. Sumário das estratégias de análise funcional de tabelas.	28
Tabela 8. Exemplo extraído de Hurst (2000), pp.85, para ilustração de análise estrutural.	28
Tabela 9. Exemplo extraído de Hurst (2000) para ilustração do objectivo do seu método.	30
Tabela 10. As empresas da amostra (a firma foi escrita, quando possível, como surge no relatório).....	36
Tabela 11. Tabela de frequências dos diferentes tipos de linhas.	44
Tabela 12. Medidas da distribuição dos espaços interiores nos diferentes relatórios.....	48
Tabela 13. Simulação do erro quadrado médio para diferentes valores de γ	55
Tabela 14. Exemplo de tabela que implica a opção por um k elevado (mínimo 9).	56
Tabela 15. Exemplo de tabela que implica a opção por um i elevado (mínimo 5).....	56
Tabela 16. Regra de construção dos parâmetros do algoritmo	57
Tabela 17. Cálculo dos elementos necessários à determinação das medidas de avaliação.	62
Tabela 18. Custos associados aos erros de classificação dos registos.	64

Tabela 19. Cálculo dos elementos necessários à determinação das medidas de avaliação.	64
Tabela 20. Tipologia dos erros cometidos pelo método 5.	67
Tabela 21. Manipulação dos parâmetros do algoritmo para correcção dos problemas da EDP2000.	72
Tabela 22. Sensibilidade e economia para diversos valores dos parâmetros manipuláveis pelo utilizador.	75
Tabela 23. Avaliação do método em exemplos independentes.	76
Tabela 24. Exemplo de tabela que o método não está programado para captar.	78
Tabela 25. Sumário das estratégias de localização de tabelas.	80
Tabela 26. Tabela de contingência entre Tabela e Espaços interiores	89
Tabela 27. Matriz de “dissemelhança” entre os relatórios (a azul, os relatórios que não são considerados semelhantes).....	91
Tabela 28. Tabela de contingência entre Tabela e os distintos relatórios e contas.....	93
Tabela 29. Algumas medidas de estatística descritiva da distribuição do Número de espaços interiores por linha, nos diversos relatórios e contas	110

1. Introdução

Um relatório e contas é um documento produzido por uma empresa, com uma periodicidade mínima anual, em que se descreve o desenrolar da actividade no período em causa, as principais estratégias seguidas, e um conjunto de informações de carácter contabilístico, que permitem aos leitores ter uma noção do desempenho da empresa, sobre a qual basearão as suas decisões.

O que pretendemos desenvolver nesta tese é um método que seja capaz de “ler” um relatório e contas e encontrar nele determinada informação que lhe interessa, guardando depois essa informação numa base de dados relacional. Porém, este tipo de documento tem características muito particulares: se por um lado, o tema é relativamente estrito, por outro não contém propriamente texto corrido, tendo muitas vezes um arranjo gráfico pouco habitual, repartindo o texto em um número variável de colunas por página e contendo muitas tabelas, figuras e gráficos; além disso, a maior parte da informação que buscamos é de cariz contabilístico e costuma ser transmitida através de tabelas e não de texto. Assim, estamos diante de um problema de extracção de informação com características muito particulares.

1.1. Porquê extrair informação de relatórios e contas?

A informação contabilística das empresas tem diversas utilidades, pois transmite em termos numéricos o resultado da sua evolução no período que descreve, sendo muitos os agentes económicos cujas decisões são influenciadas por essa evolução individual. O Plano Oficial de Contabilidade identifica-os como sendo “investidores, financiadores, trabalhadores, fornecedores e outros credores, Administração Pública e público em geral”. Porém, menos referida é a existência de diversos agentes económicos com interesse na informação agregada das várias empresas. É que a combinação das evoluções de todas as empresas de um país afecta a evolução do próprio país. Por outro lado, a evolução conjunta de dado sector de actividade, afecta todos os agentes que nele

operam ou com ele possam vir a relacionar-se. Os agentes com interesse nessa informação agregada incluem: fornecedores e outros credores, financiadores, investidores, associações empresariais, trabalhadores, analistas económicos, produtores de estatística e público em geral.

O modo de recolha desta informação não permite, porém grande velocidade no tratamento dos dados, porque ou as entidades que recebem informação ou as que as prestam têm de preencher questionários manualmente. Além de ser moroso e de impedir o tratamento de maiores quantidades de dados, o trabalho é desmotivante para as pessoas envolvidas.

Se se criasse um método de extrair essa informação automaticamente, aumentaria a quantidade de empresas cuja informação os agentes económicos podem tratar com os recursos que têm à sua disposição. Isto teria diversas consequências positivas: melhoraria a representatividade das amostras utilizadas; aumentaria a probabilidade de se extrair conhecimento dos dados; diminuiria o erro esperado na inferência, a partir dos dados, de descritores do Universo de referência. Assim sendo, as decisões dos agentes económicos estariam mais bem fundamentadas e as pessoas que, nas diversas entidades, preenchem manualmente o formato da base de dados poderiam dedicar-se a actividades de maior valor acrescentado, para si e para a entidade em que trabalham.

1.2. O reporte financeiro em Portugal

O decreto-lei n.º 410/89 de 21 de Novembro instituiu o Plano Oficial de Contabilidade em Portugal, que explicita os modelos de tabelas que se devem seguir para apresentação da informação contabilística. Como modelos que são, as empresas adaptam-nos às suas necessidades de transmissão de informação em dado exercício, não tendo sido invulgar, neste período de transição para o Euro, ver empresas que acrescentam novas colunas às tabelas para apresentar o valor de cada item em ambas as moedas correntes. Outro factor relativamente habitual é as empresas não apresentarem todas as rubricas do modelo senão apenas aquelas em que têm valores não nulos.

Nos países da Comunidade Europeia, as empresas com determinadas características são obrigadas por lei a depositarem os seus relatórios e contas anuais em instituições públicas que têm obrigação de os divulgar. No nosso país, essas instituições são as Conservatórias do Registo Comercial, que recebem grandes quantidades de informação, que divulgam individualmente a quem o solicitar. No entanto, grande parte da utilidade de compilar toda essa informação fica por explorar. Em Espanha, a criação de um formato único para o depósito e a utilização de um *software* de leitura óptica (OCR) permitiu que se compilasse a informação recolhida para uma base de dados, que permite a publicação de estudos económicos, bem como a publicidade de qualquer relatório através da Internet.

Por outro lado, as empresas com títulos admitidos à cotação em Bolsa de Valores têm obrigações mais regulares de divulgação de dados, sendo os seus relatórios e contas distribuídos pela Bolsa de Valores através do seu Jornal e imagens digitalizadas de seus excertos transmitidas pela Internet. Novamente aqui a informação divulgada é apenas individual. Porém, nos Estados Unidos a recolha destes dados em formato XML (eXtensible Markup Language) tornou a totalidade dos relatórios facilmente transmissíveis pela Internet (não apenas excertos), tendo sido criados métodos que permitem a extracção de informação específica destes documentos, comparável para as diversas empresas depositantes. Note-se que os princípios contabilísticos geralmente aceites naquele país (US-GAAP) não definem modelos para a construção das tabelas financeiras, havendo maior dificuldade do que haveria em Portugal em encontrarem-se conceitos compatíveis entre as diferentes empresas, principalmente a níveis mais detalhados da informação.

1.3. O reporte financeiro no futuro

A partir de 2005, as empresas europeias com títulos admitidos à cotação em Bolsa de Valores terão por obrigação publicar os seus relatórios e contas já não de acordo com os princípios contabilísticos geralmente aceites no país em que têm as suas sedes ou em que estão cotadas, mas sim de acordo com os princípios emanados nas Normas

Internacionais de Contabilidade (IAS). Tal como o US-GAAP, estas normas também não prevêem actualmente modelos para o reporte financeiro, o que significa que haverá maior diversidade na forma como as empresas dispõem e denominam a informação que devem divulgar nos seus relatórios e contas. Isto terá consequências sobre o nível de detalhe ao qual será possível comparar os conteúdos dos conceitos publicados pelas empresas. Existindo alguma sensibilização a nível europeu sobre a questão da manutenção de uma estrutura mínima pré definida de apresentação das contas, o futuro é incerto: caminhamos para a manutenção dos formatos ou para uma tendência mais anglo-saxónica? A resposta a esta questão poderá vir a exigir a adaptação do funcionamento do método para tratar os dados das empresas que estão vinculadas às IAS. A questão é tanto mais premente quanto haver a hipótese daquelas normas virem, a longo prazo, a ser aplicáveis a uma percentagem cada vez maior de empresas.

Entretanto, também são de esperar profundas transformações na capacidade de transmitir e tratar automaticamente a informação publicada pelas empresas. O XBRL (eXtensible Business Reporting Language) é uma adaptação da linguagem XML ao contexto específico em que operamos. Nesta nova linguagem foram definidas etiquetas (*tags*) que identificam univocamente todos os itens que devem aparecer num relatório e contas. Estas etiquetas devem estar adaptadas aos princípios contabilísticos geralmente aceites a que o relatório obedeça, existindo já dicionários de etiquetas (taxonomias) criados ou em processo de criação para o US_GAAP, para as normas internacionais de contabilidade e para o Código Comercial Alemão, entre outros.

A aplicação generalizada desta linguagem resolveria muitos problemas que esta tese se propõe resolver, e com maior grau de segurança. Porém, ainda se podem contar alguns anos até que tal aconteça e a necessidade de tratar e utilizar dados agregados de empresas é de hoje. Para além disso, as diversas fases deste método poderão ser directamente aplicadas ou facilmente adaptadas a outros domínios, que não sofrerão tal uniformização brevemente, como o caso, por exemplo, de informação apresentada em tabelas em textos científicos ou jornais ou mesmo páginas Web.

No capítulo dois, explicaremos em que consiste a extracção de informação, visitaremos os âmbitos nos quais a comunidade científica se debruçou sobre o estudo de tabelas e abordaremos as propostas de alguns autores para resolver as principais tarefas associadas ao estudo e utilização computacional de tabelas: localização, segmentação, análise funcional, análise estrutural e interpretação. No terceiro capítulo, apresentaremos sucintamente os dados com que iremos trabalhar e as características genéricas que gostaríamos que o método de extracção de informação tivesse, dedicando o resto do capítulo à descrição e avaliação de um processo de localização de tabelas inseridas em texto. O apêndice 1 apresenta testes estatísticos efectuados para suportar algumas das decisões implícitas no método de localização descrito, contendo os anexos 1 e 2 um conjunto de medidas de Estatística descritiva analisadas para o mesmo efeito. Os anexos 3 e 4 resumem os principais resultados obtidos na fase de avaliação do método de localização.

Por fim, no capítulo quatro planearemos os passos necessários para que um sistema, que terá por *input* um qualquer documento, tenha por *output* um conjunto de dados específicos extraídos das suas tabelas. Isto passará pela definição de estratégias de resolução das diversas tarefas ligadas ao estudo de tabelas, que deverão ser cuidadosamente integradas para permitir uma utilização óptima da informação (implícita ou explícita) que o documento contém e uma máxima segurança no *output* que se produz. O Apêndice dois apresenta um algoritmo que se aplicará a uma dessas tarefas.

2. Extracção de informação de tabelas contidas em texto

2.1. O que é extracção de informação?

A primeira pergunta que é suscitada pelo título desta tese (e deste capítulo) é “O que é extracção de informação (*information extraction*)?”. “O objectivo da Extracção de Informação é obter dados específicos a partir de um documento”, Riloff e Lehnert (1994), isto é, “o *input* do sistema é um linha de texto e o *output* é alguma representação da informação a ser obtida desse texto”, Hobbs e Israel (1994). A definição dada pelo EAGLES Lexicon Interest Group (1998) é bastante clara: “Os sistemas de extracção de informação extraem dada informação pré-definida acerca de entidades e das relações entre elas a partir de um texto em linguagem natural e colocam esta informação num registo estruturado de uma base de dados ou num *template*”. Neste contexto, o *template* funciona como um formulário, a preencher automaticamente com a informação contida em vários documentos, um por documento, sendo os *templates* de diferentes documentos comparáveis entre si.

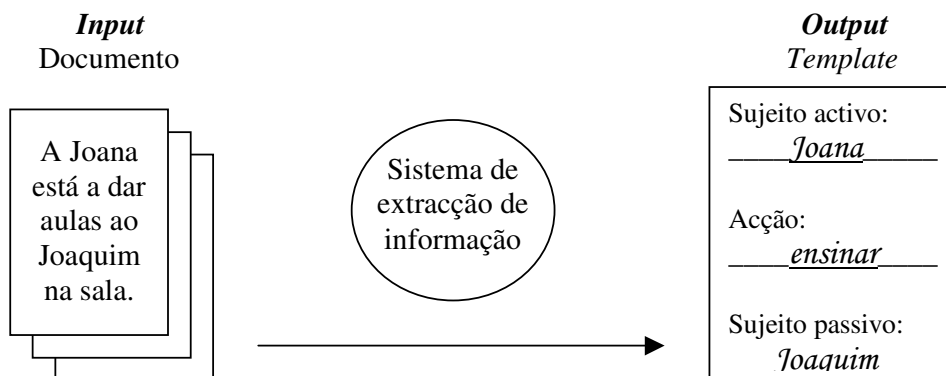


Figura 1. Um exemplo de problema de extracção de informação.

A Extracção de Informação é um dos campos de uma área de estudos mais abrangente – o Processamento de Linguagem Natural, que estuda os idiomas humanos a partir de uma perspectiva computacional. Outros campos desta área são, por exemplo, a

Extracção de Conhecimentos de Texto (*Text Mining*), que tem por objectivo “a descoberta, reconhecimento e derivação de nova informação a partir de um grande conjunto de textos”, Chang *et al.* (2000), e a Busca de Informação (*Information Retrieval*), cujo objectivo é a obtenção de “informação relevante a partir de um amplo conjunto de textos, sendo as suas técnicas tipicamente utilizadas para obter documentos relevantes a partir de um amplo conjunto de vários tipos de documentos”, Constantino (1997), entre outras.

2.2. O estudo de tabelas

“A table is an object which uses linear visual cues to simultaneously describe logical connections between the discrete content entries in the table. A content entry is the basic component of information in a table”, which “can be any visual symbol”.

Cameron (1989)

2.2.1. Áreas de investigação que se interessam por tabelas

De acordo com Hurst (2000), as tabelas são elemento de estudo de quatro grandes áreas:

- Edição e formatação: que está ligado ao modo de formatar uma tabela de modo a convir da melhor forma possível a informação que pretende transmitir. Sendo um problema já antigo, data de 1962 um dos primeiros livros que trata de tabelas num ambiente computacional.
- Psicolinguística: trata do “desenho, integração e utilização de tabelas em termos dos processos humanos envolvidos”, lidando particularmente com “o modo como as tabelas são lidas e o efeito que certos princípios organizacionais terão sobre a velocidade de localização de uma informação na tabela”.
- Análise da estrutura de documentos: regularmente o ponto de partida desta área de estudo é um documento em *bitmap*, mas também pode ser texto regular, e o que se pretende é encontrar métodos para descrever a sua estrutura, física e lógica (um

interessante exemplo de tal descrição é apresentado no ponto 4 de Klink et al. (2000)). Em relação às tabelas, pretende “localizá-las, segmentá-las e analisá-las”, computacionalmente (Hurst, 2000).

- Obtenção de conhecimentos específicos a partir de documentos, como Busca e Extracção de informação, que tradicionalmente operam a partir de texto. O interesse deste campo pelas tabelas advém de se ter observado que não poucas vezes uma tabela encerra o conteúdo de todo um texto e, enquanto se continuasse a servir apenas de texto corrido, estaria a desperdiçar-se um manancial de informação.

Em boa verdade, é muitas vezes difícil definir a fronteira entre aqueles últimos dois campos, isto porque:

1º) a obtenção de conhecimento específico a partir de tabelas pressupõe resolvido o problema anterior, o da sua localização, segmentação e análise, tarefas que os investigadores terão de resolver quando não partem de documentos devidamente etiquetados (*marked-up*);

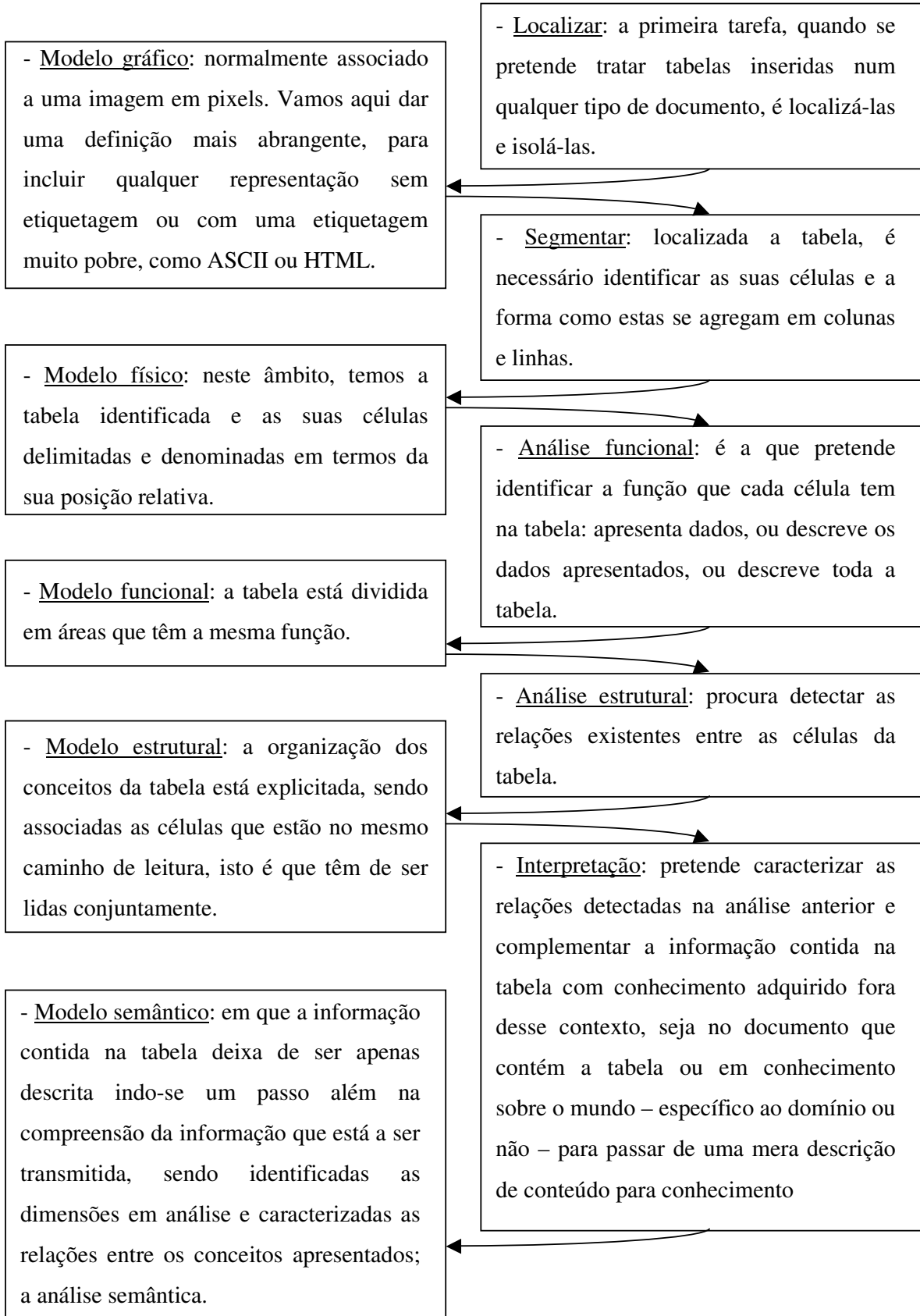
2º) a análise de documentos tem por vezes como *output* o conteúdo das células que surgem nas tabelas, expressando-o em termos de relações hierárquicas.

2.2.2. Modelos de representação de tabelas e tarefas associadas

Existem, de acordo com Hurst (2000), cinco grandes modelos pelos quais se pode representar uma tabela, cada um mais informativo que o anterior. Estes diferentes modelos servem de *input* e *output* às diferentes tarefas associadas ao estudo de tabelas, as quais já fomos referindo: localização, reconhecimento, análise funcional, análise estrutural e interpretação. Na página seguinte, explicam-se as diferenças entre os modelos e os objectivos das tarefas, de um modo que se espera esclareça a relação entre uns e outras.

Modelos de representação de tabelas

Tarefas associadas ao estudo de tabelas



Os estudos dos autores que trabalham nestas áreas podem ser enquadrados ao longo da cadeia de tarefas / modelos de representação, tal como se pode visualizar na Figura 2.

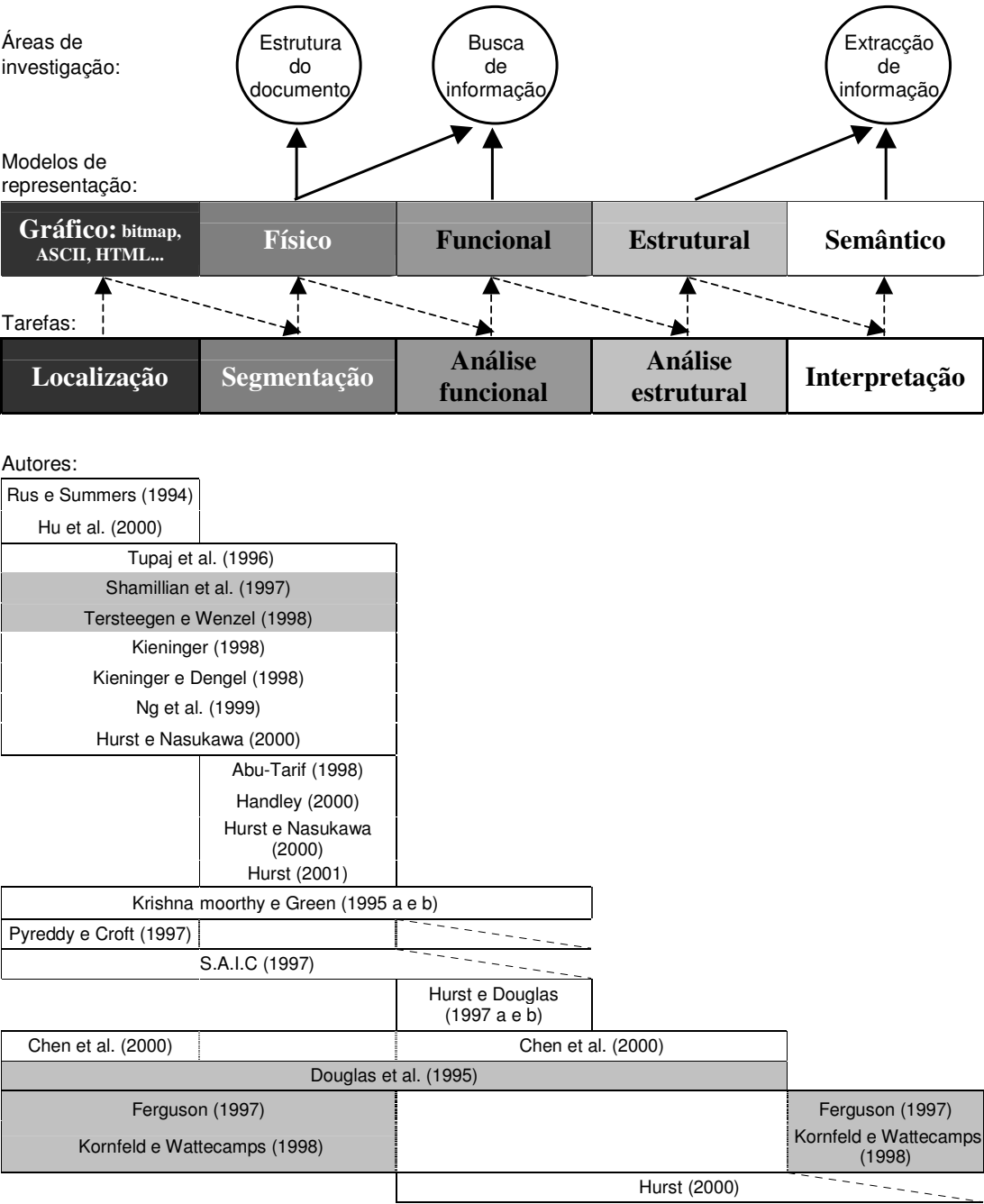


Figura 2. Relações genéricas de *input-output* entre as tarefas e modelos de representação associados às tabelas.
Posicionamento de autores seleccionados (a cinzento, os estudos mais dependentes do contexto).

A análise de tabelas tem ganhado cada vez mais espaço na cena da investigação a nível mundial, sendo já comum verem-se espaços dedicados a ela em conferências científicas, como sendo: Document Analysis Systems (DAS) 2002, que se propunha lançar um desafio sobre tabelas, que entretanto foi cancelado, e em que um dos temas centrais foi “*Tables and forms*”, no qual foi eleito o melhor *Student paper*; o *workshop* Web Document Analysis (WDA) 2001 teve uma segunda apresentação com o título “*Web as 2D Documents: Tables and Images*”; e a Document Recognition and Retrieval VIII, que teve uma sessão dedicada ao tema “*Tables Understanding*”.

2.3. Tarefas na análise de tabelas

2.3.1. Localização

Sempre que se analisa um determinado documento em que estão inseridos diversos elementos gráficos e em que temos por objectivo específico as tabelas, o primeiro passo a dar é localizar as tabelas dentro do documento, o que passa por, como diz Tupaj et al. (1996) “diferenciá-las de outros elementos como o corpo do texto, os cabeçalhos, os títulos, as bibliografias, as listas, [...] os desenhos, as imagens” e os gráficos. Mesmo em linguagens formais que têm etiquetas (*tags*) para identificar tabelas, como é o caso do HTML, o problema coloca-se, pois várias vezes tais etiquetas são utilizados para facilitar um determinado aspecto gráfico a estruturas que nada têm a ver com tabelas, surgindo por vezes também tabelas não assinaladas. Os métodos escolhidos para resolver o problema, que descrevemos nos pontos seguintes, dependem obviamente do formato do documento de origem.

2.3.1.1. Em ficheiros de imagem

Rus e Summers (1994) consideram que um bloco de texto é tabela se for possível delimitar-lhe não apenas linhas, mas também colunas, e se os conteúdos das células destas forem similares entre si. Assim, calculam, para cada bloco de cada página, a percentagem de pontos brancos por linha vertical. Dado que o número de pixels brancos numa linha vertical segue distribuição binomial (em que a probabilidade de um pixel ser

branco é obtida a partir da dimensão média das palavras escritas num texto em inglês), se aquela percentagem estiver acima da média mais h desvios padrão, há uma probabilidade $1 / h^2$ de ser um separador de coluna, sendo h definido pelo utilizador. Para cada coluna, o conteúdo de cada célula é analisado com o objectivo de definir se são parecidos entre si ou se são de tipos semânticos muito distintos, utilizando-se depois um algoritmo especializado para decidir se estamos diante de uma tabela.

Green e Krishnamoorthy (1995b) criaram um método em que o utilizador começa por fornecer ao sistema um conjunto de características base das tabelas que deseja ver tratadas, i.e., um modelo. Para detectar tais tabelas num documento, o sistema procura os efeitos visuais que o seu modelo diz rodearem a tabela, que podem ser linhas ou espaços, seleccionando como tabelas todas as áreas por eles envolvidas.

Tupaj et al. (1996) começam por procurar os delimitadores verticais e horizontais que organizam o texto na página, analisando a distribuição de pixels negros por linha horizontal (e depois vertical). A existência de delimitadores verticais significa que o conteúdo da página está distribuído em duas ou três colunas. Se estas colunas forem de texto corrido, serão amplas e estarão delimitadas por relativamente pequenos corredores verticais de espaços. Cada segmento rectangular assim definido é então testado para verificar se tem características de tabela ou não, sendo os que têm apresentados a um *software* de leitura óptica (OCR). Cada linha de cada bloco é depois classificada como sendo vazia, ou contendo uma única expressão (palavra ou número) rodeada de quatro ou mais espaços, ou contendo várias expressões. Seguidamente, procuram-se no texto palavras chave que podem aparecer no título, como por exemplo “tabela”, e daí para baixo (e depois para cima, se necessário) procuram-se determinadas sequências de linhas que se consideram típicas de uma tabela, marcando-as como tal até se encontrarem quatro linhas seguidas vazias, considerando-se aí que se chegou ao fim da tabela. A avaliação do método foi feita em dois ambientes distintos: em jornais científicos, obteve precisão (*precision*)¹ de 93%, recuperação (*recall*)² de 92.4% e acerto (*accuracy*)³ de 93%; em relatórios e contas de empresas americanas, obteve

¹ Precisão = total de tabelas correctamente detectadas / total de tabelas detectadas.

² Recuperação = total de tabelas correctamente detectadas / total de tabelas existentes.

³ Acerto = $2 * \text{Recuperação} * \text{Precisão} / (\text{Recuperação} + \text{Precisão})$.

precisão 71,4%, recuperação 75% e acerto 71%. O autor afirma que este segundo ambiente tem maior diversidade de tabelas, justificando assim os desempenhos inferiores. Como desvantagens do método os autores afirmam não detectar colunas que tenham um mesmo cabeçalho e criar falsas colunas, i.e., maioritariamente vazias.

Shamillian et al. (1997) criaram uma interface com o qual o utilizador pode descrever o aspecto da **tabela tipo** que o sistema deve procurar, comparando-se o padrão de pixels da projecção vertical de cada linha do texto com o de cada linha da tabela modelo, escolhendo-se a linha modelo a que maximiza a semelhança com a linha de texto, isto é, em que mais vezes se encontram o mesmo tipo de pixel em determinadas posições. Se este máximo, depois de normalizado, estiver abaixo de dado limite, considera-se que não se trata de uma linha da tabela alvo. A diferença entre este método e o de Green e Krishnamoorthy (1995) é que este serve para tabelas com formato fixo em que os campos têm comprimentos pré-definidos, e aquele é utilizado para tabelas cujo formato varia algo, em número de colunas, por exemplo.

Hu et al. (1999) desenvolveram um método aplicável a documentos disposto numa só coluna (como está esta tese), quer em texto simples quer em imagem. Para todas as possíveis delimitações de início e de fim de uma tabela numa página, calcula-se o seu “mérito” como tabelas como a soma ponderada das correlações entre as respectivas linhas (que é dada pela média ponderada de duas medidas: o número de pares de espaços em branco que têm a mesma posição nas duas linhas subtraídos dos que não a têm; e 2,5 vezes o número de *word bounding boxes* (palavras, se em ASCII) que têm dimensão similar e se “sobrepoem”, subtraídas do número das que estão assim dispostas). São consideradas tabelas as linhas que maximizam o mérito acima de dado limite. A avaliação foi feita em mensagens de *e-mail* e 25 imagens digitalizadas páginas de jornal; em imagem, a precisão foi de 81%, a recuperação de 91%, e o acerto de 81%; em texto, obteve-se precisão 83%, recuperação 93% e acerto 83%.

2.3.1.2. Em texto simples

Douglas et al. (1995) calculam, para todos os segmentos de espaços brancos, a sua posição e se o texto contíguo é numérico ou não. Depois, agrupam-se as linhas

contíguas que são atravessadas por um corredor vertical de espaços brancos (ou *plumb line*) e segmenta-se horizontalmente o texto nos pontos em que se encontram duas ou mais linhas vazias. Uma heurística é utilizada para caracterizar cada bloco assim definido como sendo uma coluna de tabela: deverá ser estreita e delimitada por espaços brancos relativamente grandes. O passo seguinte é consolidar estas colunas, medindo a semelhança entre elas: se semelhantes, à maior parte dos segmentos de uma deverá corresponder segmentos noutra. As áreas semelhantes são unidas, sendo aplicada uma heurística às colunas destas áreas para definir quais são tabelas de facto.

Pyreddi e Croft (1997) constróem um histograma com o número de caracteres verticalmente alinhados em k linhas contíguas, com k igual a 3. Considera-se que os intervalos em que esse número se encontra sempre abaixo de dado limite são separadores de coluna. Se, num conjunto de linhas contíguas, houver três ou mais separadores, temos uma tabela. O teste de avaliação no Wall Street Journal obteve precisão 98,2%, recuperação 94,8% e acerto 98,2% na classificação de cada linha.

A Science Applications International Corporation (1997) criou um programa que converte automaticamente documentos estruturados em SGML (Standard Generalized Markup Language). A cada linha do documento, o método aplica um conjunto de heurísticas, para a classificar como “não é o início de tabela” ou “foi encontrada toda a tabela que aqui começa”, observando-se tantas linhas subsequentes quanto necessário.

Kienenger (1998) criou um sistema que não procura delimitadores de colunas, mas sim o seu complementar, porções de texto agrupáveis num conjunto. A partir de uma palavra (ou *word bounding boxes* (wbb) se em imagem), vão-se agregando todas as que com ela estão verticalmente alinhadas, até se terem percorrido todas as palavras de cada grupo de linhas contíguas, isolando-se assim blocos de texto. Quando um destes blocos contém apenas uma palavra por linha, considera-se que se trata de uma coluna de tabela.

Ng et al. (1999) utilizou o algoritmo C4.5 de indução de árvores de decisão, aliado a um algoritmo de *back propagation* para redes neuronais artificiais, para gerar um classificador que defina se dada linha pertence ou não a uma tabela. São utilizados 27

atributos, sobre a linha actual, a precedente e a seguinte: se a linha é vazia; o número de espaços que a indenta; a posição do primeiro caracter; a do último; se a linha contém caracteres diferentes de número, letra ou espaço; o número de segmentos com dois ou mais espaços contíguos; *idem*, com três ou mais espaços contíguos; *idem* com dois ou mais separadores contíguos (. , * e -); *idem* com três ou mais separadores contíguos. O teste no Wall Street Journal conseguiu um acerto de 95%, em validação cruzada.

2.3.1.3. Em ficheiros etiquetados

Chen et al. (2000) definem um sistema simples para localizar tabelas em HTML, HyperText Markup Language: 1º. são extraídas as etiquetas que identificam as tabelas; 2º. são excluídas as tabelas que não tenham pelo menos duas células (um atributo e um valor) ou as que tenham demasiadas hiper ligações ou figuras; 3.º partindo do princípio que valores de um mesmo atributo são parecidos, calculam-se três medidas de semelhança entre cada célula e as suas vizinhas - o número de caracteres em comum, se pertencem ou não à mesma categoria semântica (como sendo datas, percentagens, unidades monetárias, nomes próprios) e se são ou não ambas numéricas. Se a percentagem de células semelhantes estiver acima de dado limite, considera-se que se trata de uma tabela. O teste sobre informação de voos no Yahoo chinês conseguiu precisão e recuperação de 92,9% e 80,1%, respectivamente, e acerto de 86,0%, obtendo valores mais elevados quando a categoria semântica é utilizada na medida de semelhança entre as células.

Ferguson (1997) e Kornfeld e Wattecamps (1998) procuram tabelas específicas em ficheiros texto que incluem algumas etiquetas (*tags*) de XML, eXtensible Markup Language. Assim, o início de uma tabela está identificado com a etiqueta TABLE, o seu fim com a *tag* \TABLE, os seus títulos com a *tag* CAPTION e cada uma das suas colunas com a etiqueta C, reservando-se a *tag* S para marcar as colunas de referência. Ferguson (1997) utiliza um conjunto de atributos das linhas, como seja o número de colunas, para encontrar candidatas a tabelas, percorrendo depois a área acima destas à procura de expressões contabilísticas, atribuindo uma classificação aos títulos mais próximos dos que procura extrair. Para identificar os títulos, toma em consideração o tipo de letra, informação de *layout*, a proximidade a uma possível tabela, guardando o

melhor resultado. Kornfeld e Wattecamps (1998) extraem as tabelas utilizando uma versão de um *parser* LR(k)⁴ adaptado para o efeito, sendo muito lacónicos na descrição dos demais procedimentos utilizados.

Na tabela abaixo, pretendeu-se salientar algumas características dos métodos apresentados, de modo a identificar os mais comuns e os mais raramente escolhidos e de certa forma esboçar a localização de cada método / autor ao longo dos vectores identificados. Como se observa na coluna total, a maior parte dos métodos é de aplicabilidade genérica, baseia-se na distribuição de espaços no documento e utiliza heurísticas para encontrar tabelas, usando-se regularmente parâmetros pré-definidos.

		Rus e Summers	Douglas et al.	Krishnamoorthy e Green	Tupaj et al.	Shamillan	Pyreddi e Croft	Ferguson	SAIC	Kornfeld e Wattecamps	Kieninger	Hu et al	Ng et al.	Chen et al.	Total
		1994	1995	1995	1996	1997	1997	1997	1997	1998	1998	1999	1999	2000	
Informação tida em conta	Linhas desenhadas			x											1
	Distribuição de espaços	x	x	x		x	x		x				x		7
	Posição relativa das palavras				x						x				2
	Medidas de semelhança		x			x						x		x	4
	Palavras chave				x			x							2
	Análise lexical	x												x	2
Método	Busca de delimitadores	x	x	x	x								x		5
	Agrupamento de blocos										x	x			2
Classificador	Heurísticas	x	x	x	x		x	x	x		x			x	9
	Padrões				x										1
	Maximização de dada medida					x						x			3
	Modelo induzido pelos dados												x		1

⁴ Um *parser* LR é um tipo de *parser* que analisa o seu *input* da esquerda para a direita para reconhecer uma porção de texto como provinda ou não de uma linguagem especificada por uma dada gramática de livre contexto. (k) refere-se ao número de símbolos do *input* que se permite ao sistema visualizar adiante do que está a classificar para basear uma decisão. Este tipo de algoritmos tem as seguintes vantagens: pode ser adaptado a todas as linguagens de programação; pode ser implementado eficientemente; detecta erros sintácticos. (in *Wikipedia, the free encyclopedia*, www.wikipedia.com).

		Rus e Summers	Douglas et al.	Krishnamoorthy e Green	Tupaj et al.	Shamillian	Pyreddi e Croft	Ferguson	SAIC	Kornfeld e Wattecamps	Kieninger	Hu et al	Ng et al.	Chen et al.	Total
		1994	1995	1995	1996	1997	1997	1997	1997	1998	1998	1999	1999	2000	
Uso de parâmetros	Pré definidos		x		x	x	x		x					x	6
	Definidos por utilizador	x		x			x								3
Aplicabilidade do método	Genérica	x	x		x		x		x		x		x		7
	Específica			Adap tável		Adap tável		XML		XML		uma colu- na		HTM L	6

Tabela 1. Sumário das estratégias de localização de tabelas.

2.3.2. Segmentação

O objectivo mínimo desta tarefa é basicamente delimitar as células de dada tabela, definindo assim o seu modelo físico. O problema parece elementar, mas, no Apêndice A da tese de doutoramento de Hurst (2000), é apresentado um amplo conjunto de diferentes de tabelas, que ilustram o tipo de dificuldades que podem surgir num sistema que se pretenda generalista.

A Figura 3 pretende dar uma noção das relações entre os diferentes níveis de estruturas que constituem as tabelas. A verdade é que o modelo gráfico não tem geralmente mais que pixels ou caracteres, dependendo de ser imagem ou já texto; um conjunto de caracteres contíguos pode ser uma palavra ou um número, a que chamamos, na figura abaixo, genericamente de expressão; ainda antes do reconhecimento de caracteres, é possível definir com algum grau de confiança a fronteira delimitadora de cada expressão (*word bounding box*), a que chamaremos simplesmente *wbb*. Conjuntos de expressões estão geralmente inseridos em células da tabela, podendo pertencer também aos seus títulos ou rodapés. As células por sua vez agrupam-se para formar as colunas e linhas da tabela, que podem ter cabeçalhos próprios.

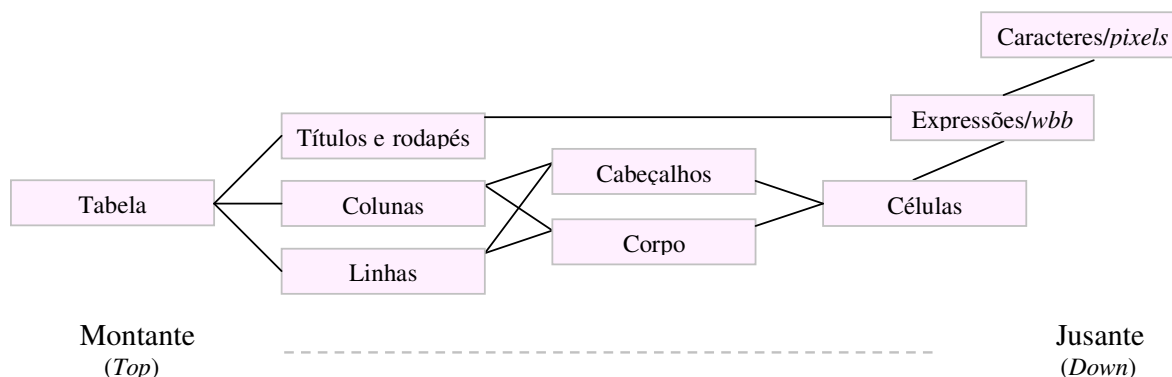


Figura 3. Estruturas que compõem o modelo físico da tabela.

A generalidade dos autores segue uma abordagem *top-down* em que, a partir da tabela como um todo, se procuram delimitadores, sejam rectas, sejam corredores de espaços brancos, sejam ambos, que definam as linhas e as colunas, assignando depois as células a estas estruturas. Mas existem também abordagens *bottom-up*, em que, se vai expandindo dada expressão ou wbb, para a unir a outras que devam pertencer à mesma estrutura – célula, coluna, linha e finalmente tabela.

Há ainda sistemas que produzem apenas uma descrição física da tabela, definindo as células e as suas posições relativas e ainda, eventualmente, a dos seus delimitadores, havendo outros que fazem por ir mais longe, pressupondo que os aspectos gráficos detectados são indícios da relação existente entre as diferentes células e procurando reflectir isso no modo de denominar as células detectadas.

Green e Krishnamoorthy (1995b) utilizam uma estratégia *top-down*, para ir segmentando a tabela em estruturas cada vez menores até chegar às células. O número de segmentações a fazer e os delimitadores que se procuram em cada uma é definido por um modelo do tipo de tabela que se está a tratar. Os delimitadores podem ser rectas, que são distinguidas em função das suas espessura e tracejado, e corredores de espaços brancos, diferenciados pela sua largura. A cada segmentação, são atribuídos nomes às novas estruturas, compostos por uma letra, que muda quando a divisão é vertical e um número, que muda quando é horizontal, sendo a denominação de cada célula o resultado da concatenação das denominações das estruturas hierarquicamente superiores.

Kirby Puckett's Hits 1987-1989	
Year	Hits
1987	207
1988	234
1989	215

(a) Tabela original

Kirby Puckett's Hits 1987-1989	
Year	Hits
1987	207
1988	234
1989	215

(b) Delimitadores encontrados

A1A1A1	
A1A2A1	A1A2B1
A2A1A1	A2A1B1
A2A2A1	A2A2B1
A2A3A1	A2A3B1

(c) Denominação das células em função das estruturas detectadas

Tabela 2. Exemplo extraído de Green e Krishnamoorthy (1995a) para ilustração do seu método.

No passo final do processo de localização de tabelas, Douglas et al. (1995) tinham já delimitado as colunas das suas tabelas, mas faltava delimitar as linhas: se a tabela tiver linhas vazias, estas delimitam as linhas da tabela; caso contrário, será necessário agrupar as linhas físicas para encontrar as linhas da tabela, método que não é explicado.

Para segmentar as colunas das tabelas previamente localizadas, Tupaj et al. (1996) procuram corredores de espaços verticais que atravessem partes significativas da tabela, tomando por delimitadores de colunas o seu ponto médio, medido na linha em que o corredor é mais estreito (ver o exemplo na tabela anexa). Se uma determinada linha interromper um número significativo de corredores verticais, considera-se que é um título ou um rodapé da tabela. Os autores apontam as seguintes desvantagens do seu sistema: não agrupa células que abarcam várias linhas e podem-se formar colunas fantasma, apontando como solução possível para este ponto uma análise semântica. A avaliação feita em jornais científicos e em relatórios e contas alcançou respectivamente precisão de 97% e de 92% e recuperação de 91.6% e de 75%.

Exemplo		
Salários pagos pela empresa XPTO, SA		
Secção	2000	1999
Armazém	5.000.000	4.500.000
Peixaria	250.000	230.000
Padaria	50.000	29.000
Caixas	3.000.000	2.700.000
Total	8.500.000	7.659.000

Pintam-se a azul os corredores verticais de espaços que delimitam as células; considera-se que a primeira linha não pertence à tabela por interromper um número significativo de corredores verticais; o delimitador das duas primeiras colunas é definido como sendo igual ao seu ponto médio medido na linha "Armazém", aonde o corredor é mais estreito.

Tabela 3. Exemplo para ilustração de Tupaj et al. (1996).

Shamillian (1997) simplesmente utiliza o padrão da projecção vertical da linha modelo para delimitar cada célula da linha em análise. Por seu lado, Ferguson (1997) analisa os cabeçalhos das colunas para definir os seus limites aproximados, extraíndo depois cada linha, até detectar o fim da tabela – os cabeçalhos das colunas e o fim da tabela estão identificados em XML.

As linhas de tabelas seleccionadas por Kornfeld e Wattecamps (1998) na fase anterior são cortadas em células pelos espaços que separam as expressões, tal que a cada linha corresponda pelo menos uma célula de texto, não mais numerosas que as colunas da tabela. Um algoritmo seguidamente amalgama as células de texto quando os seus conteúdos, por serem demasiado largos para o espaço disponível, abarcam várias linhas.

Um dos problemas de pós processamento que Kieninger (1998) tem de resolver é o de isolar as colunas de tabelas em que uma das células abarca mais de uma coluna: serão isoladas numa só coluna todas as palavras em alinhamento directo com apenas uma expressão na linha de baixo. Um segundo problema consiste em ter de agregar as células que abarcam diversas linhas. Para tal, Kieninger distingue entre dois tipos de blocos: os que são constituídos por apenas uma expressão por linha, denominados tipo 1, e os que têm mais de uma palavra, chamados tipo 2. Cada bloco de tipo 1 é individualizado na

sua própria célula; se ao lado de blocos tipo 1 houver blocos de tipo 2, então a segmentação destes obedece à segmentação daqueles, agrupando-se as células abarcadoras, desde que a primeira palavra da linha seguinte não caiba no espaço ainda

<u>Exemplo</u>		
Secção	Salários pagos pela empresa	
	2000	1999
Armazém de congelação	5.250.000	4.730.000
Padaria	250.000	229.000
Caixas	3.000.000	2.700.000
total	8.500.000	7.659.000

Primeiro problema: na fase anterior, as colunas com números são inicialmente amalgamadas numa só, por estarem em alinhamento vertical directo com a expressão “Salários pagos pela empresa”. Serão isoladas numa só coluna as expressões numéricas que estejam em alinhamento vertical directo com apenas uma expressão na linha seguinte e vice-versa.

Segundo problema: “Armazém de Congelamento” ocupa duas células em vez de apenas uma. O método detectará que a expressão está inserida num bloco de tipo 2, contíguo a um bloco de tipo 1 e segmentará as células de ambos de acordo com as linhas detectadas no bloco de tipo 1, verificando que “congelamento” não cabe no espaço disponível na célula da linha anterior.

Tabela 4. Exemplo para ilustração de Kieninger (1998).

disponível da célula actual. Kieninger e Dengel (1998) adicionaram a este sistema um módulo que permite expressar o *output* em HTML.

Ng et al. (1999) distingue três tipos de caracteres: alfanuméricos {números, letras}, espaços e restantes, considerando que a mudança de tipo de caracteres ao longo de uma linha é um indicador da demarcação das colunas da tabela. Assim, para encontrar os delimitadores destas, cada linha vertical da tabela é caracterizada com seis atributos, três que resultam da comparação com a linha vertical anterior e três com a seguinte: a percentagem de vezes em que há transição de alfanumérico para espaço ou restante; *idem* de espaço ou restante para alfanumérico; *idem* para manutenção espaço ou restante. Estes dados são classificados por uma árvore de decisão gerada em C4.5 com *back propagation* por redes neuronais. Por outro lado, considera-se que, dado que diferentes linhas da tabela apresentam informação similar, se duas linhas são parecidas, elas são diferentes linhas da tabela. Desta vez são considerados apenas dois tipos de carácter, espaço ou não, apresentando-se quatro atributos ao classificador, três que medem a percentagem de transições de e para espaço ou a manutenção de espaço em comparação com a linha anterior e um que dá a posição do primeiro carácter não espaço na linha actual. Os classificadores foram gerados pelos mesmos algoritmos que os das colunas. Os testes no Wall Street Journal, obtido por validação cruzada, conseguiram acertos de 90% e 84% nas colunas e linhas respectivamente.

O sistema de Handley (2000) determina a estrutura da tabela a partir de imagem, dando como *output* uma lista de células, as expressões que contêm, as coordenadas dos seus delimitadores, da sua linha e da sua coluna. Este autor afirma que “como é típico em muitos sistemas de reconhecimento de documentos que operam em ambientes diversos, um desempenho adequado depende de sequências integradas de regras cuidadosamente desenhadas e menos de modelos analíticos”. Assim sendo, ele começa por detectar se a tabela está ou não totalmente delimitada por rectas. Caso não esteja, utiliza um conjunto de heurísticas, com alguns parâmetros, para, utilizando como *input* a localização das rectas, os espaços brancos que separam as *word bounding boxes* e a posição relativa destas, para ir refinando as diversas estimativas do seu *output*.

Hurst (2001) utiliza um método perfeitamente novo para detecção de blocos de texto (objectivo similar ao de Kieninger (1998)), utilizando primeiro informação de tipo espacial, depois refinada com informação linguística. Num primeiro passo, o sistema detecta pontos em que há que decidir se o texto prossegue na mesma linha ou se na margem esquerda da linha seguinte: entre palavras separadas por um corredor vertical de espaços; ou quando o espaço

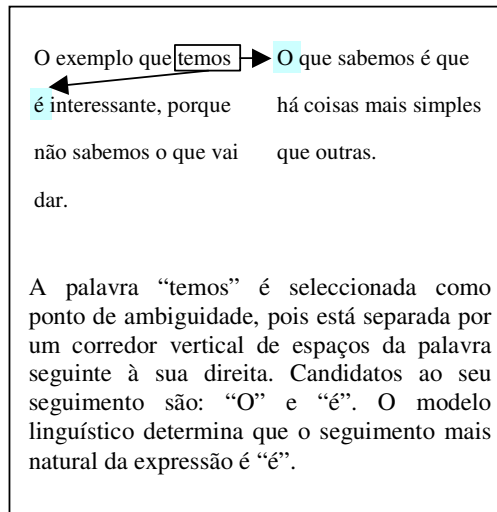


Figura 4. Exemplo para ilustração de Hurst (2001).

entre duas palavras é grande; ou na margem direita da folha. Para estes pontos, o seguimento do texto pode estar em três localizações distintas: ser a próxima palavra à direita; ou a primeira palavra da linha seguinte; ou qualquer dos pontos da linha seguinte que estão à esquerda da posição actual. O sistema terá então de decidir, entre os candidatos, qual deles é a extensão linguisticamente mais coerente do segmento em causa, dentro do modelo da linguagem em que o texto está escrito. Se uma palavra tem vários antecedentes possíveis, escolhe-se aquele que é mais provável. Estas decisões baseiam-se num modelo de linguagem, que contém o número de vezes que duas palavras apareceram juntas num texto de treino que lhe serviu de base. A avaliação do método foi efectuada em relatórios e contas de empresas americanas (SEC *fillings*) e, recorrendo-se a um modelo de linguagem específico do contexto, conseguiu-se um desempenho em tabelas de 99,4% e uma recuperação de 78,9%.

Em resumo e como fica claro na tabela seguinte, a maior parte dos métodos baseia-se em heurísticas, procurando encontrar como delimitadores espaços brancos ou rectas, havendo alguns que se baseiam em agrupamentos de blocos a partir da posição relativa das palavras. O único método que detecta hierarquias entre variáveis é o que se baseia

num modelo pré-definido do tipo de tabela que se está a procurar. O uso de parâmetros continua também a ser muito comum.

		Green e Krishnamoorthy	Douglas et al.	Tupaj et al.	Shamillan	Ferguson	Kornfeld e Watticams	Kieninger	Ng et al.	Handley	Hurst	Total
		1995	1995	1996	1997	1997	1998	1998	1999	2000	2001	
Informação tida em conta	Rectas (<i>line work</i>)	x								x		3
	Espaços brancos	x	x	x		x					x	5
	Posição relativa das palavras/wbb				x			x		x	x	3
	Semelhança entre células				x				x			1
	Mudanças de tipologia de caracteres						x		x			1
	Coerência linguística										x	1
Método	Busca de delimitadores (<i>top-down</i>)											
	- Heurísticas		x	x		x	x			x		4
	- Modelo induzidos pelos dados								x			1
	- Modelo inserido pelo utilizador	x			x							1
	Agrupamento de blocos (<i>bottom-up</i>)							x			x	2
Aplicabilidade do método	Genérica		x	x				x	x	x	x	7
	Específica	Adap-tável										1
Uso de parâmetros pré definidos			x	x						x		4

Tabela 5. Sumário das estratégias de reconhecimento de tabelas.

2.3.3. Análise funcional

O objectivo desta análise é o de agrupar as células da tabela em subconjuntos que cumprem a mesma função, por exemplo, distinguindo nomeadamente entre as células que contêm dados e as que os descrevem.

O sistema de Douglas et al. (1995) opera no contexto específico da indústria da construção. Para identificar os domínios das tabelas aplicam um conjunto de heurísticas,

que têm em conta as coordenadas de cada célula, o tipo semântico do seu conteúdo na sub-linguagem do domínio (exemplo, “Portland” é um tipo de cimento) e um booleano que identifica se existe uma relação semântica entre o conteúdo da célula e o título da tabela.

Green e Krishnamoorthy (1995) criaram um sistema que tem por base um modelo, parametrizável, do tipo de tabela que se está a tratar e que “fornece as características essenciais que a tabela exibirá se for reconhecida e coordena o mapeamento das estruturas lógicas a partir dos atributos gráficos”. Do mesmo modo que foi o modelo que definiu o modo de segmentar a tabela nas suas estruturas constitutivas, é ele que liga estas aos seus significados. Por exemplo, relativamente à Tabela 2, o modelo informaria que A1A1 ? é o título da tabela e A1A2 ? são os cabeçalhos das colunas.

Pyreddy e Croft (1997) têm por objectivo caracterizar as linhas da tabela como contendo atributos ou valores, que permita uma adequada busca (*retrieval*) de tabelas. Para esse efeito, usam um conjunto de heurísticas baseadas na largura dos espaços interiores à linha (grandes indicam valores), no alinhamento (contém um atributo uma linha no início ou no fim da tabela ou uma linha no centro que tenha um alinhamento de espaços brancos diferente da média), na dimensão (se a dimensão de uma linha for diferente das restantes, será atributo), em padrões e ainda no número de colunas de cada linha (se as primeiras linhas têm menos colunas que a média, são de atributos). Se uma linha se segue a uma de atributos e formato diferente do da média da tabela, é sujeita ainda às seguintes regras: atributos são texto, enquanto valores são números; atributos estão em maiúsculas e valores em ambas. O teste no Wall Street Journal conseguiu precisão de 90,6%, recuperação de 81,4% e acerto de 85,7%.

Tendo por objectivo converter automaticamente documentos para SGML, o sistema implementado pela SAIC não vai muito longe na análise funcional. Como diz Thompson (1996), “a etiquetagem de tabelas contém muita informação sobre o aspecto da tabela... mas muito pouca sobre como a tabela relaciona as células”. A heurística utilizada pela Science Applications International Corporation (1997) para distinguir os cabeçalhos das colunas é simplesmente que estarão na primeira linha da tabela que

tenha todas as colunas excepto talvez a primeira preenchidas, o que será um dos motivos pelos quais eles admitem não conseguir lidar com tabelas com células que abarcam várias colunas.

Hurst e Douglas (1997 a) (1997 b) definiram um modelo simples de tabela e utilizam um conjunto de medidas de coesão entre as células de diferentes áreas rectangulares da tabela para avaliar até que ponto dada área pode ser uma instanciação de uma das alternativas do modelo teórico. (A medida de coesão depende da comparação do número de caracteres numéricos e alfabéticos de cada duas células e das suas dimensões e é calculada como a média ponderada de uma estimativa da probabilidade de uma célula ser descritora de outra e da semelhança entre as células). Com as porções rectangulares da tabela que ultrapassam um dado limite de coesão, um algoritmo procura construir um *puzzle* que cubra toda a tabela e em que não haja sobreposições. A avaliação do método atribuiu-lhe pouca capacidade de distinguir descritores de valores.

Hurst (2000) fala nos seguintes métodos para alcançar o objectivo de caracterizar uma célula como pertencente a um conjunto dos atributos ou a dados ou a nenhum deles:

- heurísticas, por exemplo: as áreas de dada tipologia de células devem ser rectangulares; todas as células sob ou ao lado de uma célula que abarca mais que uma linha ou coluna devem ter a mesma tipologia entre si; se todas as células filhas de uma célula abarcadora forem de referência, aquela também o deve ser; se uma palavra é repetida numa tabela, todas as células que a contenham devem ser do mesmo tipo; *strings* repetidas são células de referência; células que abarcam todas as colunas da tabela são de referência; as margens esquerda ou de topo têm células de referência. Além disso, menciona também: regras semântico-sintácticas, de tipo autómatos finitos, que indicam a classificação a atribuir a determinados tipos semânticos, quando encontrados em dado arranjo físico; regras de conteúdo, que assumem que os conteúdos das células de referência da tabela são discutidos no texto em frases que se referem à tabela ou nos títulos. Aquando do uso de uma combinação de heurísticas, a avaliação por validação cruzada em documentos provindos de diversos contextos obteve precisão

e recuperação (ponderadas pela dimensão das classes) de 92.8% e 25.0%, respectivamente.

- modelos bayseanos, com atributos físicos, como o número de células adjacentes à que está a ser classificada, e com atributos de conteúdo - de comparação entre o conteúdo de cada célula e as palavras encontradas no documento. Além disso, dá-se especial atenção aos títulos de cada secção do texto (claramente identificado no *input* em questão, SGML), às frases que contenham a palavra “tabela” e ao título da tabela – qualquer palavra destas frases que esteja na tabela implica que a sua célula será de referência. O teste, no mesmo conjunto de documentos, obteve precisão e recuperação, ponderadas pela dimensão das classes, de 91.1% e 91,4% respectivamente, tendo-se concluído que a utilização de atributos de conteúdo aumenta ligeiramente a qualidade dos resultados.

- identificação de padrões, que passa por três fases – 1. definição de padrões físico em termos de dada célula (no caso, inspeccionam-se as células que se cruzam com perpendiculares à célula em causa nas quatro direcções, até uma profundidade d pré determinada); 2. comparação dos padrões com novos exemplos e 3. para lidar com o conjunto de padrões que correspondem a um exemplo em particular, há que contar e eventualmente ponderar o número de votos para cada classificação, havendo tantos votos quantos padrões aplicáveis. O teste, no mesmo conjunto de documentos, obteve precisão e recuperação, ponderadas pela dimensão das classes, de 92,3% e 90.3% respectivamente, tendo-se concluído que: a profundidade 2 foi melhor que o *naive bayes* para a classificação das células de referência; o método melhora com maior profundidade; o método de votação por maioria foi o melhor dos testados.

- a combinação dos três métodos atinge máximas recuperação e precisão: 93.4% e 93.1%, respectivamente.

Para determinar se a primeira linha contém atributos, Chen et al. (2000) medem a semelhança⁵ de cada linha com a última linha da tabela: se o número de linhas similares for maior que $(n-2)/2$, em que n é o número total de linhas da tabela, e a semelhança

⁵ Os atributos que servem à medição da semelhança, identificados na secção 2.3.1.3. são: o número de caracteres em comum, se pertencem ou não à mesma categoria semântica (como sendo datas, percentagens, unidades monetárias, nomes próprios) e se são ambas numéricas ou não.

entre a primeira e a última linha for menor que a das suas congêneres, então a primeira linha contém atributos. Um método análogo determinará se a primeira coluna contém atributos. Caso existam células que abarcam diversas linhas ou colunas, como é o caso da Tabela 6(a), a tabela será dividida em sub-tabelas, delimitadas pelas células abarcadoras, sendo aquela avaliação feita de modo independente em cada sub-tabela (ver Tabela 6(b)). Note-se que a classificação em atributo ou valor é feita para linhas e colunas e não para células individuais.

Código de viagem			DP9LAX01AB	
Validade			99.04.01-00.03.31	
Classe			Económica	Extensão
Adulto	PREÇO	Individual	34.450	2.510
		Duplo	32.500	1.430
		Cama extra	30.550	720
Criança		Quarto	25.800	1430
		Cama extra	23.850	720

(a) tabela original

Código de viagem	Código de viagem	Código de viagem	DP9LAX01AB	DP9LAX01AB
Validade	Validade	Validade	99.04.01-00.03.31	99.04.01-00.03.31
Classe	Classe	Classe	Económica	Extensão
Adulto	PREÇO	Individual	34.450	2.510
Adulto	PREÇO	Duplo	32.500	1.430
Adulto	PREÇO	Cama extra	30.550	720
Criança	PREÇO	Quarto	25.800	1430
Criança	PREÇO	Cama extra	23.850	720

(b) as células abarcadoras foram replicadas pelas células que abarcam e as três sub-tabelas pintadas a cores diferentes: a primeira corresponde à junção da área branca a azul claro; a segunda, à junção da área azul claro e escuro e a terceira à junção da área azul escura e cinzenta.

Tabela 6. Exemplo extraído de Chen et al. (2000) para ilustração do funcionamento do seu método.

A tabela seguinte caracteriza os métodos vistos em termos de algumas características. Como podemos verificar, as heurísticas continuam a ser o método mais utilizado para classificar cada célula quanto à sua tipologia, havendo bastante diversidade nos indicadores utilizados.

		Douglas et al.	Green e Krishnamoorthy	Pyreddi e Croft	SAIC	Hurst e Douglas	Hurst	Chen	Total
		1995	1995	1997	1997	1997	2000	2000	
Informação tida em conta	Posição relativa das células	x	x				x		3
	Espaços brancos			x					1
	Alinhamento			x					1
	Dimensão relativa do conteúdo			x		x			2
	Categoria semântica do conteúdo	x					x	x	3
	Comparação do conteúdo com o texto em redor	x					x		2
	Número de colunas preenchidas por linha			x	x				2
	Tipologia de caracteres					x			1
	Semelhança entre células			x		x		x	3
Classificador	Heurísticas	x		x	x		x	x	5
	Comparação com modelos teóricos de tabelas		x			x			2
	Padrões			x			x		2
	Modelo formal						x		1
Uso de parâmetros pré definidos						x	x		2

Tabela 7. Sumário das estratégias de análise funcional de tabelas.

2.3.4. Análise estrutural

Hurst (2000) define esta tarefa como a de “para cada célula da tabela, determinar o conjunto de células que podem ser alcançadas a partir dela navegando na tabela”. Navegar na tabela é percorrer-lhe as células que estão ligadas entre si, cuja leitura tem de ser feita conjuntamente, pois

Exemplo			
Cidade	Assassinatos		Variação percentual
	1990	1996	
Nova Iorque	2.245	984	-56%
Los Angeles	983	688	-30
Chicago	854	791	-7
Houston	568	261	-54
Filadélfia	503	431	-14

Tabela 8. Exemplo extraído de Hurst (2000), pp. 85, para ilustração de análise estrutural.

os respectivos conteúdos complementam-se. Assim, por exemplo, na tabela da figura ao lado, as células pintadas a azul devem ser lidas conjuntamente, encontrando-se as cinco no mesmo caminho de leitura. Aliás, navegando pela totalidade das células da tabela, poder-se-ia preencher uma tabela relacional com a informação ali contida (a qual poderia ser facilmente utilizada para extracção de informação).

A leitura da tabela em Chen et al. (2000) tem por objectivo formar pares atributo / valor o que será feito de modo independente em cada uma das sub-tabelas delimitadas pelas células abarcadoras. Se, na fase anterior, não se tiverem identificado quaisquer colunas ou linhas de referência para a sub-tabela, cada célula será emparelhada com a primeira célula da sua coluna. Se se tiver concluído que a sub-tabela contém uma linha e uma coluna de atributos, serão formados, para cada dado, dois pares. Assim, por exemplo na Tabela 6b), para 30.550 seriam constituídos dois pares: {Cama extra; 30.550} e {Económica; 30.550}. Dos diversos pares atributo / valor assim formados sobressaem duas características: uma célula pode ser valor de mais de um atributo, caso em que os atributos deverão ser unidos num só, por exemplo {Cama extra; Económica; 30.550}; e algumas células são atributo numa relação e valor noutra, sobressaindo uma hierarquia de atributos a que se pode aplicar transitividade: dos pares {Adulto; Preço}, {Preço; Cama extra} e {Cama extra; Económica; 30.550} resulta {Adulto; Preço; Cama extra; Económica; 30.550}. Desvantagem: uma tal hierarquia de atributos só é detectada caso existam células abarcadoras.

Hurst (2000) utiliza um conjunto de heurísticas para identificar o caminho de leitura da tabela. Estas entram em conta não só com a posição física das células e a sua classificação em referência ou dados, feita no passo anterior, como também com redes semânticas que se constróem em torno dos conteúdos. São utilizados módulos independentes para identificar relações entre diferentes células de referência (em que se investiga as possíveis ligações entre uma célula de referência e outras que se localizem acima ou ao lado dela) – e entre dados e referências (em que simplesmente se une cada célula de dados à primeira célula de referência acima e à esquerda dela). O teste, para o mesmo conjunto de documentos da fase anterior e o mesmo método de avaliação, conseguiu uma precisão e recuperação de 98,1% e 91,5% respectivamente.

2.3.5. Interpretação

Até aqui, localizou-se a tabela e segmentaram-se-lhe as células, as quais foram caracterizadas e depois ligadas entre si por se crer que existe uma relação entre elas que obriga a que sejam lidas em conjunto. Porém, a relação existente entre os conteúdos das células, nomeadamente entre as células de referência, não foi caracterizada, não se sabendo ainda nada sobre aspectos qualitativos ligados ao porquê da escolha de determinada organização das células na tabela.

2.3.5.1. Obtenção das categorias em análise

Hurst (2000) descreve um método interessante que permite identificar as diferentes modalidades referidas na tabela para um dado atributo. Na figura seguinte, apresenta-se, a título de exemplo, a tabela utilizada pelo autor para expor o método (página 104 da sua tese) e as categorias identificadas para a mesma.

<i>States</i>		E	B
Q	<i>Sequence</i>	Q	QQ
	<i>Probability</i>	1.0	.2
R	<i>Sequence</i>	R	QR
	<i>Probability</i>	0.0	.1

Categorias identificadas:

{*States.Q*, *States.R*}

{*Sequence*, *Probability*}

{E,B}

Tabela 9. Exemplo extraído de Hurst (2000) para ilustração do objectivo do seu método.

Sumariando o método, a partir do modelo estrutural definido no passo anterior, Hurst (2000) **1)** reconstitui o caminho de leitura completo de cada célula de dados, passando por todas as células que estão unidas até encontrar o limite superior e à esquerda da tabela. Seguidamente, **2)** calcula as dependências, isto é, identifica-se o conjunto de células de referência que aparecem sempre e apenas na presença de outras células do mesmo tipo, computando o número de vezes em que cada associação acontece e respeitando a restrição que nenhuma célula pode aparecer no conjunto das dependências

mais vezes do que o que aparece no conjunto anterior. Como é provável que surjam várias dependências para dada célula, preferem-se as que estão alinhadas e, quando alinhadas à esquerda e em cima, as mais comuns. Sendo a dependência transitiva, **3)** define o conjunto de células maximamente dependentes aplicando aquela propriedade ao conjunto anterior. A partir deste novo conjunto, **4)** identifica as categorias, como sendo os conteúdos das células que, no conjunto anterior, nunca aparecem juntas. Consideram-se disjuntas as categorias que nunca aparecem juntas nos caminhos de leitura da tabela, sendo conjuntas as que aparecem associadas. A precisão e recuperação medidas no teste ao método foram de 93.2% e 95.0%, respectivamente.

2.3.5.2. Caracterização das relações entre células

Sabemos agora quais são os atributos apresentados na tabela e que valores podem apresentar, sendo já possível desenhar com bastante precisão uma árvore de hierarquia dos conceitos apresentados. Porém, falta caracterizar as relações detectadas. Como diz Hurst (2000), “interpretar e compreender a tabela, depois de se terem estabelecido as categorias, pode ser caracterizado como procurar a informação em falta [...] que pode ser descoberta no documento propriamente dito, em modelos do domínio do qual deriva o documento ou em conhecimento universal.”

Este autor prevê as seguintes classificações para as relações existentes entre células:

1. super-tipo nominal: algumas células contêm concretizações ou qualificações do conteúdo de outra (exemplo: País – Portugal; Carro – Verde);
2. quantitativa – unidade de medida: tipo de relação existente entre uma célula que contenha uma unidade de medida e as restantes (exemplo: 5.000.000 – contos);
3. quantitativa – valor do atributo: tipo de relação existente entre uma célula que contém um dado e a que contém a descrição desse dado (exemplo: Despesas de instalação – 100);
4. partitiva: o conceito explicitado em dada célula está contido no de outra (exemplo: Activo, Imobilizações corpóreas; cozinha: fogão).

Além daquelas, prevêem-se duas classificações linguísticas para o conteúdo de dada célula: disjuntiva – toda a descrição da célula está contida nela própria; ou elíptica – caso em que o conteúdo de uma célula deva ser lido em conjunto com o de outra, como por exemplo, se uma célula disser “Número de pessoas interrogadas” e outra diz “Número de mulheres”.

Hurst (2000) descreve, no seu Apêndice F, uma pequena gramática que lhe permite identificar pedaços de texto como sendo datas ou unidades de medida ou unidades monetárias ou anos, entre outros grupos semânticos de interesse, a partir do modo como são escritos, explicitando as diferentes alternativas. O seu sistema compara os conteúdos de dada categoria da tabela com a gramática, de modo a caracterizá-la. Seguidamente, heurísticas são utilizadas para decidir qual o tipo de relação de que se trata. Além disso, a presença nos descritores de algumas palavras chave, como sendo “*type*” ou “*member*” desencadeia a que se assuma uma relação de super-tipo nominal. A rede semântica construída a partir dos itens de uma categoria também é utilizada. O autor admite que os métodos propostos estão ainda em fase embrionária, não fazendo qualquer teste.

2.3.5.3. Interpretação em contexto específico

Kornfeld e Wattecamps (1998) e Ferguson (1997) construíram sistemas similares entre si para interpretar os relatórios e contas de empresas cotadas nos Estados Unidos – os denominados SEC *fillings*.

Num contexto específico, eles puderam servir-se das restrições que o contexto impõe sobre os dados como ferramenta para o processo de interpretação, tendo Ferguson (1997) identificado as seguintes: existem requisitos que definem que informação tem de estar presente nas tabelas, a qual está definida por padrões de *strings* facilmente detectáveis; os números estão somados em subtotais – se as contas não acertarem, ou a extracção foi mal feita ou há erro na informação de base; certas regras contabilísticas têm de estar presentes, como por exemplo Activo = Passivo + Capital Próprio. Em Portugal, todas estas restrições também são aplicáveis.

Assim, eles detectam as relações partitivas dos elementos apresentados no balanço e na demonstração de resultados utilizando como indicadores as relações matemáticas existentes entre as partes, cuja soma aparece num todo que figura no documento, e que, por sua vez, é uma parcela de um todo mais abrangente. Kornfeld e Wattecamps (1998) servem-se igualmente da indentação para detectar a hierarquia presente nos conceitos apresentados. Estes autores consideram que conseguem tratar convenientemente uma tabela quando chegam a um total final que é o resultado da soma de todas as parcelas da tabela e que está nela escrito.

Com esta estrutura, constróem uma árvore de hierarquias dos conceitos apresentados, que preenchem com os descritores de cada linha. Cada elemento da árvore será comparado, palavra a palavra, com uma colecção, construída manualmente, de alternativas possíveis para nomear dada variável de interesse. Nesta comparação, começa-se pelo topo da árvore, indo-se gradualmente descendo até às folhas, sendo que a classificação dos itens mais acima restringe o espaço de busca dos itens mais abaixo – o que não só é computacionalmente mais económico como fundamental no contexto contabilístico, em que expressões muito similares podem ter posições diferentes na estrutura de conceitos.

Ferguson (1997) menciona adicionalmente algumas técnicas utilizadas para assegurar a fiabilidade da extracção: sempre que possível, compara dados provenientes de diferentes tabelas, quando se referem ao mesmo conceito; verifica que certos rácios, construídos com as variáveis extraídas, tenham valores razoáveis; identifica a localização inicial da informação extraída e facilita-lhe o acesso; permite que o sistema corrija um erro numa coluna utilizando números de outra coluna; infere a unidade monetária utilizada em dada tabela, sempre que esta não esteja identificada, por comparação com a escala de outras tabelas aonde tal tenha sido feito; utiliza um sistema de *rankings* para as diferentes soluções perseguidas pelo modelo, para que ele possa seguir a segunda melhor solução se a primeira claramente não servir.

Quanto ao tratamento das notas do relatório e contas, este autor começa por identificar o estilo utilizado no primeiro título que encontra para a mesma, caracterizando o modo

como este título está expresso, que será regular ao longo do documento, usando depois esta caracterização para encontrar os títulos das restantes notas, os quais costumam estar enumerados sequencialmente.

Para avaliação do seu método, Ferguson (1997) utilizou 212 *filings*, localizando 94.3% das demonstrações de resultados e 93.8% dos balanços e obtendo precisão e recuperação de 90% ou mais na extracção da maioria dos itens individuais. Kornfeld e Wattecamps (1998), por sua vez, dizem ser capazes de tratar eficazmente e na totalidade 85% dos relatórios e contas, devendo-se os restantes 15% a erros aritméticos da informação de base ou a irregularidades de formatação, mas sendo relativamente simples de corrigir por um humano, quando postos lado a lado o documento de base e a tabela já parcialmente somada.

3. Um método de localização de tabelas em texto

O problema que nos propusemos resolver no início desta tese implica encontrar soluções para resolução de cada uma das tarefas apresentadas no capítulo anterior (ilustradas na figura 2), num tipo específico de documento – os relatórios e contas que as empresas, por lei, publicam numa periodicidade mínima anual. O objectivo é extrair destes documentos os dados de algumas tabelas contabilísticas, que a partir daqui chamaremos genericamente de tabelas interessantes (às restantes tabelas contidas no documento, chamaremos normais), e que são as seguintes:

- . Balanço
- . Demonstração de resultados por natureza
- . Demonstração de resultados por função
- . Demonstração de fluxos de caixa
- . Empresas do grupo, associadas ou participadas
- . Movimentos ocorridos nas contas do activo imobilizado
- . Movimentos ocorridos nas amortizações e provisões do activo imobilizado
- . Movimentos ocorridos nas contas de provisões
- . Movimentos ocorridos nas rubricas de capital próprio
- . Demonstração do custo das mercadorias vendidas e matérias consumidas
- . Demonstração da variação de produção
- . Demonstração de resultados financeiros
- . Demonstração de resultados extraordinários

Neste capítulo, começaremos por apresentar sucintamente o nosso conjunto de treino, i.e., o conjunto de relatórios que nos serviram de exemplo para derivar o método proposto, descrevendo a seguir algumas das dificuldades a que o método de extracção de informação deverá responder de modo robusto e algumas das características genéricas que gostaríamos que o sistema tivesse. O resto do capítulo descreve um processo de localização de tabelas inseridas em texto: as variáveis que podem apoiar a decisão, o classificador criado e as suas variantes, os testes de avaliação e a opção pela variante com melhor desempenho, terminando-se com a identificação de aspectos específicos desta tarefa que poderão vir a ser aprofundados.

3.1. A selecção da amostra de relatórios e contas a tratar

Obtivemos uma amostra de relatórios e contas publicados por empresas portuguesas nos seus sites de Internet. Na tabela seguinte, apresentamos a lista das empresas e dos exercícios a que se referem os relatórios conseguidos.

<i>ID</i>	<i>Empresa</i>	<i>Exercício</i>	<i>N.º de páginas</i>
1 a 3	Auto-Industrial, S.A.	1998 a 2000	34, 43, 56
4 a 5	BRISA - Auto-estradas de Portugal, S.A.	1999 e 2000	101, 87
6	Cimentos de Portugal, SGPS, S.A. (Grupo Cimpor)	2000	234
7 a 10	Grupo EDP	1997 a 2000	24, 22, 169, 196
11	Grupo Galpenergia	2000	107
12	Lusomundo - Sociedade Gestora de Participações Sociais, S.A.	1999	33 (de 70)
13	Novabase - S.G.P.S., SA	2000	12
-	ParaRede	2000	48
14	Portucel - Empresa Produtora de Pasta de Papel, SA	2000	103
15	RAR - Sociedade de Controle (Holding), S.A.	2000	32 (de ?)
16	Shell Portuguesa, Lda.	1999	41
17	Siemens, S.A.	2000	38 (de 70)
18 a 19	Sonae, SGPS, S.A.	1999 e 2000	81, 79

Tabela 10. As empresas da amostra (a firma foi escrita, quando possível, como surge no relatório).

Evidentemente, este não é um conjunto representativo do Universo de relatórios e contas existentes em Portugal, pois a amostra contém apenas empresas de razoável dimensão, com nomes bem conhecidos do público, várias das quais cotadas em Bolsa de Valores, que utilizam este instrumento como um modo de darem a conhecer a sua empresa e que detêm equipas de marketing que se encarregam da sua produção, para que cada página seja o mais agradável à vista possível.

Isto é diferente da realidade de muitas das nossas pequenas e médias empresas, que, ou vêm neste instrumento apenas um modo de cumprir uma obrigação legal de divulgação de informação, ou detêm uma equipa demasiado pequena para poderem dispensar o tempo necessário à produção de um relatório visualmente diverso.

Também a nossa amostra de exercícios económicos não é minimamente representativa de exercícios económicos normais, pois entre 1999 e 2001 as empresas foram-se gradualmente adaptando à entrada em vigor do Euro, sendo comum encontrarmos, principalmente em 2000, relatórios expressos em duas moedas: escudos (ou contos) e euros: algumas empresas optaram por apresentar todas as peças contabilísticas em duplicado, um exemplar de formato habitual por cada moeda, porém outras optaram por acrescentar em cada tabela uma ou duas colunas com os valores na outra moeda.

Porém, estes relatórios pertencem a um subconjunto de empresas / períodos que à partida introduzem maiores factores de complexidade em termos de tratamento automático. Ora, uma vez que não nos é possível obter uma amostra representativa dos relatórios existentes em Portugal, parece-nos importante procurar incluir nos exemplos de treino que utilizemos um conjunto que, tanto quanto podemos avaliar, contenha maior dificuldade que o regular, para que o método possa ser mais generalizável. Além disso, temos relatórios de dimensões muito diversas e de sectores de actividade distintos (considerando que não estamos interessados em relatórios de empresas financeiras).

De uma análise manual prévia feita aos exemplos, constatou-se o seguinte: a ParaRede terá de ser excluída, pois apresenta o balanço e demonstração de resultados como imagens; por seu lado, os anos de 1997 e 1998 da EDP apresentam balanços extremamente sintéticos; Brisa 1999 tem um erro grave e extremamente raro na apresentação do balanço (uma coluna saltou para a página seguinte); e à Autoindustrial em 1999, falta-lhe a primeira parte do balanço. Assim, estes exemplos terão de ser excluídos do conjunto de treino na fase de interpretação, sendo interessante guardá-los para teste, para ver o tipo de alertas que o método gerará. O documento da Novabase não é um verdadeiro relatório, mas antes uma “Comunicação de facto relevante”, que ainda assim contém um balanço e uma demonstração de resultados, pelo que mantê-lo-emos na amostra, como um factor para aumentar a capacidade de generalização do modelo que criemos. Por fim, os documentos da Lusomundo, RAR e Siemens incluem apenas a parte do relatório em que se publicam as contas; a versão completa era constituída por 7, 9 e 3 partes respectivamente.

Os relatórios publicados na Internet estão geralmente em formato Acrobat (pdf), que passamos para texto simples (ASCII) com a ajuda de um programa utilitário distribuído gratuitamente (pdftotxt). O apêndice 1 mostra alguns exemplos do aspecto original destes documento e do seu correspondente em txt.

Depois de passados a ASCII, os dezanove relatórios são importados para uma base de dados Access, cujos registos são as linhas do relatório. Para cada uma, identifica-se o número da respectiva página (PAGINA), e, relativamente a esta, o número de ordem da linha (LINHA), a qual é também identificada em relação ao total do documento de origem (ID). Depois, atribui-se um número também a cada uma das linhas não vazias do relatório (IDC). A variável ESQUERDA capta a indentação do conteúdo da linha e é obtida como a diferença entre o comprimento da *string* original e o comprimento da *string* que resta após se eliminarem todos os espaços à esquerda. Por fim, substituem-se na *string* todos os pares de espaços contíguos por um só até não haver mais candidatos à substituição. Subtraindo o comprimento da *string* assim obtida ao comprimento medido no passo anterior, ficamos com o Número de ESPAÇOS contíguos interiores à linha. LARGURA é o comprimento da *string* que resta.

3.2. Breve caracterização do arranjo do espaço em relatórios e contas

No capítulo 2, mencionámos alguns autores que experimentaram os seus sistemas em relatórios e contas de empresas americanas. De um modo geral, eles consideraram este contexto particularmente complexo, quando comparado com textos científicos ou jornalísticos, dada a elevada variedade de estilos que se utilizam.

A verdade é que estamos diante de textos escritos ora em duas colunas ora numa só que vale por duas, ora chegada à esquerda da folha ora à direita, ora ao centro (ver Figura 5); ora com gráficos e tabelas que ocupam toda a extensão horizontal da página, ora que a partilham com colunas de texto (ver Figura 6); em que as tabelas ora têm muitas colunas ora têm poucas, ora têm todas as linhas preenchidas, ora estão “interrompidas” por até dez linhas contíguas vazias (ver Figura 7); com linhas em que apenas a coluna

Órgãos sociais

Conselho de Administração

Beltrino de Azevedo, Álvaro Pereira, Paulo Jordão, Paulo Augusto,* Álvaro Paquetim*

Agente

David Bica, Edgar P. Ribeiro, J. Marcos dos Santos

Controlo de Gestão

Beltrino Mendes de Azevedo

Assessoria Fiscal

Álvaro Paquetim

Assessoria e Gestão de Risco

Álvaro Paquetim

Assessoria Legal

Beltrino Mendes de Azevedo

Gestão de Comunicação e Imagem

Álvaro Paquetim

Carta do Presidente

O DECRETO DE 1999, LAMTO DO MEDIO NA INTERPRETAÇÃO mais mediana, volta a ser novamente um ato de crescimento sustentado, que de ponto de vista técnico, quer geografico.

Os termos significativos, fundamentalmente a decisão de manter sob a designação CSM, os activos anteriormente detidos separadamente pelo Sme Investimentos, figura a figura. É a actualização do modelo de gestão financeira, com um Conselho de Administração que deverá passar a lidar com os seus elementos a propósito da generalização do modelo de gestão a todos os pontos da rede.

De ponto de vista estratégico, voltamos para um modelo de gestão de portofólio, centrado por dois sub-entidades, cada uma com a sua própria freguesia, em registos homogêneos, a uma única SMI-Holding representada por fundações de participação - através de algumas fundações e operações - em empresas de actividade em actividades, de serviços e em novos sectores. Esta forma de participação e pela soberania na perspectiva do seu potencial impacto nos registos internos de CSM e nos mercados de Fintech, SMI e Bercia e da América Latina.

As duas sub-entidades programam com elevado impacto de gestão os seus estratégias, com a ideia de garantir uma fruição nos respectivos registos e de eficácia resultante de um elevado poder de atracção de recursos humanos e financeiros, que tem permitido um significativo crescimento sustentado, uma expansão geográfica actualizada e desenvolvida, com grande planeamento e determinação, de nova medida de crescimento.

Actualmente, a SMI-HOLDING actua como uma representação de linha e registos sustentada sobre a gestão de recursos, visando uma expansão a montante com os seus investimentos em sectores chave em Espanha e sobre os Portugal e Alemanha, juntamente com outros sectores, com o apoio da SMI e da Bercia e da América Latina.

[illegible][illegible]

39

Demonstração dos Resultados por Naturezas para
o exercício findo em 31 de Dezembro de 2000
(Montantes expressos em contos)

		2000
CUSTOS E PERDAS		
Custo das mercadorias vendidas e das matérias consumidas		44 255 252
Fornecimentos e serviços externos		18 725 777
Custos com o pessoal:		
Remunerações	6 133 057	
Encargos sociais:		
Pensões	561 803	
Outros	2 283 763	8 978 623
Amortizações do imobilizado corpóreo e incorpóreo	12 796 568	
Provisões	11 744	12 808 312
Impostos	63 869	
Outros custos operacionais	55 151	119 020
	(A)	84 886 984

Figura 8. Detalhes da Portucel 2000 - células de referência abarcando várias linhas *versus* hierarquia de conceitos.

Dificuldades acrescentadas pela passagem para texto simples

Na Figura 6a), podemos observar que as tabelas foram escritas num tamanho de letra inferior ao normal de modo a caberem em apenas metade da página. Quando passado para txt, este jogo com as dimensões das letras tem as seguintes consequências (ver barras amarelas da Figura 9): no lado com letra maior, são inseridas meias linhas vazias; no lado com letra menor, por um lado teremos *strings* com dimensão muito inferior ao normal e, por outro, perder-se-á o corredor de espaços brancos que separa as diferentes colunas da página.

Uma dificuldade acrescida com que devemos ser capazes de lidar é um erro que ocorre devido à dificuldade do programa de conversão de pdf para txt em lidar com determinados tipos de letra. A Figura 10 apresenta um exemplo em que tal é notório. Evidentemente, se tal acontecer em texto corrido, não é grave para o nosso sistema, que pretendemos extraia informação de tabelas, porém quando Custo das mercadorias vendidas e matérias consumidas é transformado em “Custo das mer c a d . vendidas e das mat. c o n s u m i d a s :”, a situação poderia ser grave, pelo que o método que utilizemos para identificar um item de informação com o que pretendemos extrair deve ser robusto a erros ortográficos. De nossa parte, por exemplo, já encontrámos o título “Balano”, em vez de balanço.

(b) Esta rubrica refere-se a serviços prestados à Briser no âmbito de assistência técnica e operacional a projectos rodoviários de outras concessionárias e assistência administrativa.			Donativos Dividas incobráveis Perdas em imobilizações Correcções relativas a exercicios anteriores Outros custos e perdas extraordinários
45. Demonstrações dos resultados financeiros			Resultados extraordinários
Os resultados financeiros dos exercicios findos em 31 de Dezembro de 2000 e 1999 tinham a seguinte composição:			Proveitos e ganhos:
			Ganhos em imobilizações
			Benefícios e penalidades contratuais
			Reduções de provisões
			Correcções relativas a exercicios anteriores
			Outros proveitos e ganhos extraordinários
	2000	1999	
Custos e perdas:			
Juros suportados	8 657 566	8 029 827	
Perdas em empresas do grupo			
Diferenças de câmbio desfavoráveis	158 202	125 315	
Outros custos e perdas financeiros	1 208 802	663 215	
	10 747 901	8 838 552	
RESULTADOS FINANCEIROS	19 881 092	19 420 177	
	886 809	368 177	
Proveitos e ganhos:			
Juros obtidos	146 756	144 369	
Ganhos em empresas do grupo e associadas (Nota 10)	154 643	87 259	
Diferenças de câmbio favoráveis	585 402	122 357	
Outros proveitos e ganhos financeiros	8	14 192	
	886 809	368 177	
A rubrica de "Juros suportados" respeita a juros sobre empréstimos obtidos para o financiamento da constru-			Saldos devedores: Imposto sobre o Rendimento das Pessoas Colectivas Retenções na fonte do exercicio de 2000

Figura 9. Passagem para texto do conteúdo da Figura 5. Detalhe da Sonae 1999 – texto anormalmente distribuído por colunas. (a).

5.2. Comunicação e Relações Exteriores

No âmbito das suas actividades de comunicação entre todas as suas actividades, Comunidade em geral, e especificadamente Autoridades, parceiros de negócio e empregados, a Shell Portugal versão portuguesa do relatório social do Grupo Royal Dutch Shell 1999- O homem, o planeta e o lucro - uma atitude Este relatório, publicado pela 2ª vez consecutiva, e distribuído às autoridades e privadas, aborda a visão do Grupo Shell sobre Sustentável, sobre o enquadramento que as decisões de negócio explicitando e clarificando princípios e metas a que nos propomos, estimulando a explicitação de questões críticas ou meras

5.2. Comunicação e Relações Exteriores

No âmbito das suas actividades de comunicação entre todas as suas actividades, **Comunidade** em **geral**, e especificadamente **Autoridades**, parceiros de negócio e empregados, a Shell Portugal versão portuguesa do relatório social do Grupo Royal Dutch Shell 1999- O homem, o planeta e o lucro - uma atitude Este relatório, publicado pela 2ª vez consecutiva, e distribuído às autoridades e privadas, aborda a visão do Grupo Shell sobre Sustentável, sobre o enquadramento que as decisões de negócio explicitando e clarificando princípios e metas a que nos propomos, **estimulando** a explicitação de questões críticas ou meras

Figura 10. Detalhe da Shell 1999 – tipo de letra transposta com erro para ASCII.

Uma última dificuldade que pode surgir na fase de interpretação, já não devida à passagem de txt para pdf, é erros na informação original. Kornfeld e Wattecamp (1998) mencionam que a ocorrência de erros aritméticos em 2% dos relatórios e contas depositados na Securities Exchange Commission nos Estados Unidos os impedem de utilizar as relações matemáticas entre os conceitos como indicador de hierarquia.

3.3. Características desejáveis num sistema que trate relatórios e contas

Dissemos há pouco que o sistema tem de ser suficientemente robusto para poder lidar com erros na informação de base ou no resultado da passagem para ASCII. Mas, o que torna um sistema robusto? Vamos estabelecer que é um sistema que é capaz de:

- 1) conhecer os pontos em que é mais provável cometerem-se erros;
- 2) basear as suas decisões em métodos que sejam menos afectados por esses erros;
- 3) medir a confiança dos resultados a que vai chegando;
- 4) e registar as decisões tomadas em cada passo do processo, de modo a que possa, se a confiança de dado resultado for baixa, em *backtracking*, verificar se a confiança seria mais elevada se uma decisão passada tivesse sido diferente.

Por outro lado, é necessário que o sistema seja suficientemente rápido, pois este tipo de documentos é produzido pela maioria das empresas do país em períodos concentrados de tempo (entre Maio e Julho) e a sua análise só é útil se atempada.

Além disso, queremos um sistema interactivo. Dado que o custo do erro é elevado, queremos que o sistema questione o utilizador, enviando, por exemplo, uma mensagem para o seu correio electrónico, quando a confiança de dada decisão for baixa.

Por fim, queremos que o método seja o mais geral possível. Embora, extrair informação seja sempre um trabalho específico do seu contexto, os métodos utilizados até chegar à tarefa de interpretação não necessitam sê-lo. Assim, embora ajustemos diversos aspectos dos métodos às características das tabelas que temos por objectivo, pretendemos que o sistema seja facilmente adaptável, através da manipulação dos seus parâmetros, *templates* e restrições, a qualquer outro contexto. E mesmo na fase de interpretação, pretendemos que seja possível ao utilizador apresentar as variáveis que pretende extrair e as relações entre elas, tal que possa ser utilizável pelo sistema.

3.4. A primeira tarefa – localizar tabelas

Como já terá ficado claro, para alcançar o objectivo desta tese é necessário encontrar soluções para cada uma das fases de tratamento automático de tabelas: localização, segmentação, análise funcional, análise estrutural e interpretação. O nosso *input* é um conjunto de linhas e o nosso *output* final é o conteúdo de um pequeno conjunto de células. No processo, teremos de rejeitar várias linhas / colunas / células que não contêm a informação pretendida.

Nesta primeira fase, pretendemos, utilizando apenas a informação genérica com que se pode classificar cada linha do documento, classificar cada uma em “tem características de tabela” ou “não tem características de tabela”. Na fase seguinte, que abordaremos no capítulo 4, percorreremos, caracter a caracter, cada linha das seleccionadas e parti-las-emos nas células que a constituem, guardando estas numa nova tabela, que terá, como é evidente, uma muito maior dimensão que a tabela original. Assim, o que pretendemos neste primeiro passo, é reduzir o nosso espaço de procura para dirigir os procedimentos computacionalmente dispendiosos que se seguirão aos registos que têm, à partida, maior probabilidade de conter a informação que buscamos. Assim, não pretendemos ainda delimitar com exactidão cada tabela, apenas encontrar indícios da presença de um elemento gráfico que poderá ser uma tabela.

Num sistema em que se vão percorrer as diversas tarefas da análise de tabelas, não é necessário que cada uma seja definida de modo estanque, podendo a informação adicional recolhida em cada fase suprir de dados a fase seguinte mas também ajudar a aperfeiçoar as fases anteriores. Ora, até à interpretação, existem muitos pontos em que se podem eliminar, das linhas seleccionadas neste primeiro passo, as que foram erroneamente classificadas, ou resgatar, do conjunto das descartadas, as que o foram incorrectamente, eventualmente com menor custo que mantendo todas as linhas do documento até lá.

Há a referir que, no caso dos relatórios e contas, nem todas as tabelas são iguais: existem as que contêm informação que queremos extrair, que denominaremos

Interessantes, e as que não são o nosso objectivo, que denominaremos Normais, sendo para nós extremamente importante detectar a maior parte daquelas, mas irrelevante detectar ou não estas.

Na tabela abaixo, apresentamos os resultados da classificação manual de cada uma das 87.843 linhas dos 19 relatórios que compõem o nosso conjunto de exemplos de treino. Como se pode verificar, as linhas que nos interessam representam menos de 10% do total, sendo que menos de 20% de todas as linhas pertencem a tabelas, pelo que, se conseguíssemos distinguir *a priori*, estas das outras, teríamos um enorme potencial de economia computacional, que se repercutiria sobre as fases seguintes.

Tipologia	Frequência absoluta	Frequência relativa
Linhas de tabelas interessantes	9.685	11,0%
Linhas de tabelas normais	<u>+ 7.574</u>	<u>+ 8,6%</u>
Linhas pertencentes a tabelas	17.292	19,6%
Linhas não pertencentes a tabelas	<u>+ 70.584</u>	<u>+ 80,4%</u>
Total de linhas	87.843	100,0%

Tabela 11. Tabela de frequências dos diferentes tipos de linhas.

3.4.1. Que variáveis independentes utilizar?

3.4.1.1. Descritivas da própria linha

O problema de localizar tabelas no nosso contexto resume-se, por enquanto, a caracterizar cada linha do relatório como pertencendo ou não a uma tabela. Para atingir este objectivo, começamos por basear-nos na constatação de que as linhas com tabelas têm mais espaços interiores do que o normal no documento⁶. Para averiguarmos se esse

⁶ Foram aplicadas aos dados as técnicas de análise discriminante e a de árvores de decisão com diversas variáveis explicativas e em ambas a variável Espaços interiores foi a que demonstrou ter maior poder discriminante.

pressuposto tem algum fundamento, começámos por desenhar o histograma da variável em causa no total dos documentos de que dispomos, o qual apresentamos no Anexo 1. Daqui se retiraram duas grandes conclusões: 1. a variável apresenta indícios de ser capaz de distinguir tabelas de não tabelas; 2. não apresenta indícios de ter capacidade de distinguir as tabelas que contêm informação alvo de extracção das restantes tabelas. No ponto 1.1. do Apêndice 1, decidimos averiguar se a primeira destas conclusões tem fundamento estatístico, ficando provado que sim.

A questão agora é encontrar um método que possa ser usado para distinguir uma linha como pertencendo ou não a uma tabela com base nesta variável.

3.4.1.2. Descritivas do documento de origem

A nossa primeira questão foi saber até que ponto fazia sentido utilizar um conjunto de variáveis relacionado apenas com a própria linha e que fosse independente das características específicas de dado documento. É que, sabendo nós, como sabemos, que um relatório tem de conter um conjunto mínimo de tabelas e que todos os relatórios mais ou menos as contêm, um documento mais prolixo conterá possivelmente mais elementos que não tabelas, afectando portanto o peso dos dois grupos na amostra e logo a probabilidade de uma determinada linha ter características de tabela.

Além disso, alguns relatórios estão escritos em duas ou mais colunas de texto, separadas por corredores verticais de espaços interiores, enquanto outros terão o texto disposto, como nesta tese, numa só coluna. Ora, não poderão estas decisões de estilo afectar a relação entre o número de espaços interiores e a pertença da linha a uma tabela?

Será que, não estando os relatórios vinculados a qualquer regra de disposição da informação em cada página, não estamos a correr o risco de criar um modelo perecível a mudanças de estilo e moda ao longo do tempo, se definirmos um modelo de decisão para cada linha independente do seu contexto, principalmente quando sabemos não termos uma amostra representativa do Universo de documentos existente?

No fundo, o que queremos saber é se, por um lado, o comportamento da variável Espaços interiores é diferente de relatório para relatório, e, por outro, se a probabilidade de uma linha pertencer a uma tabela o é. A importância desta informação é que, sendo as respostas àquelas questões afirmativas, devemos afectar o modelo de decisão com características descritoras da distribuição de espaços interiores do documento de base.

No Anexo 2, demos os primeiros passos no sentido de averiguar a resposta a estas questões, usando algumas técnicas simples de estatística descritiva. Constatou-se que, entre empresas diferentes, a distribuição de espaços interiores parece ser muito distinta, variando a média entre 29 e 89 espaços e o desvio padrão entre 26 e 50; também o peso das linhas pertencentes a tabelas varia entre 6% e 37%. No Apêndice 1, resolvemos então testar estatisticamente as duas hipóteses, respectivamente nos pontos 1.2. e 1.3., ficando provado que os acontecimentos não são independentes, pelo que devemos afectar o modelo de classificação de cada linha de factores descritivos de dado relatório, nomeadamente, e dado que ficou provado que a distribuição de Espaços interiores tem capacidade de distinguir tabelas de não tabelas, de factores descritivos desta. Como fazê-lo?

Para começar, o número de espaços interiores de uma linha é o resultado da soma dos resultados de n experiências de Bernoulli, em que n é o comprimento total da linha excluídas as suas abas e em que a experiência toma valores 1 quando dada posição é ocupada por um carácter espaço contíguo a outro e 0 caso contrário. Assim sendo, podemos afirmar que o número de espaços interiores de uma linha segue distribuição binomial (assumindo válidos os testes de independência da distribuição de espaços num documento feito por Rus e Summers, 1994), sendo esta distribuição integralmente descrita pela sua média e desvio padrão teóricos (que são dados pelo produto do número de experiências pela probabilidade de sucesso, no caso da média, e disto multiplicado pela probabilidade de insucesso, no caso do desvio padrão). Assim, precisamos obter o número de vezes que se realiza a experiência, que é conhecido, e a probabilidade de ela ter resultado 1 ou 0.

Porém, ao contrário daquelas autoras, não nos parece lícito para este nosso contexto, estimar a probabilidade de um determinado carácter ser preenchido ou ser espaço a partir da dimensão média das palavras em texto português, porque relatórios e contas não são texto corrido e contêm muitos elementos cuja probabilidade de conterem espaços brancos interiores não é representada por textos normais, como acontece por exemplo em gráficos, e porque relatórios com básicas diferenças de estilo (por exemplo, com texto escritos em diversas colunas ou numa só) terão provavelmente diferentes probabilidades associadas (como ficou visto no Apêndice 1).

Uma outra alternativa seria estimarmos os valores daquela probabilidade a partir do total dos relatórios de que dispomos, através do quociente entre o total de espaços interiores e a soma do total de caracteres e espaços interiores. Porém, só se poderiam esperar resultados próximos do pretendido se detivéssemos uma amostra representativa da diversidade de relatórios existente em Portugal para os diversos períodos ou se soubéssemos pelo menos quais são e que peso têm as diferentes tipologias de relatórios neste Universo.

Assim, na falta de valores teóricos, e reconhecendo diferentes as probabilidades de dado carácter ser um espaço interior nos distintos relatórios de que dispomos, decidimos estimar a média e o desvio padrão de modo independente para cada relatório, a partir dos respectivos valores amostrais e procurar assim caracterizar a distribuição em causa.

O que quisemos ver a seguir foi se estas duas medidas de caracterização dos relatórios eram suficientes para os efeitos pretendidos ou se havia outra informação que devêssemos usar. Para isso, determinámos, para cada relatório, o limite, único e estático, que melhor separa as duas classes, usando para tal o algoritmo J48 de construção de árvores de decisão, caracterizando cada linha de cada relatório apenas pelo número de espaços brancos e a classe a que pertence. Este algoritmo encarrega-se de encontrar a primeira melhor divisória, a qual coloca na raiz da árvore. Na Tabela 12, apresentamos cada um dos relatórios e diversas medidas que se podem estabelecer para caracterizar a distribuição daquela variável bem como as divisórias do J48.

	Autoindustrial			Brisa			Cimpor	EDP			Galpenergia	Lusomundo	Novabase	Portucel	RAR	Shell	Siemens	Sonae	Sonae
ID	98	99	00	99	00	00	97	98	99	00	00	99	00	00	00	99	00	99	00
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
<i>Páginas</i>	34	43	56	101	87	234	24	44	168	196	107	33	12	103	32	41	38	81	79
<i>Média Espaços</i>	5	4	4	13	18	14	13	20	14	15	20	18	12	9	21	13	16	22	25
<i>DP espaços</i>	17	15	13	28	29	25	22	30	24	29	39	29	28	24	44	29	25	37	40
<i>Moda (excepto 0)</i>	2	2	2	6	18	14	2	62	8	8	10	6	2	4/10	2	6	10	10	16
<i>Divisória</i>	2	8	8	34	36	36	4	4	26	16	46	26	6	18	6	20	48	20	38

Tabela 12. Medidas da distribuição dos espaços interiores nos diferentes relatórios.

Aquele modelo de decisão parece indicar divisórias bastante distintas para os diversos relatórios. Decidimos pintar de cores distintas aqueles em que se aconselha um valor menor ou igual a 10, entre 11 e 20, entre 21 e 30, entre 31 e 40 e maior que 41. Podemos verificar que existem relatórios com média e desvio padrão semelhantes em que o sistema decide por divisórias muito diferentes, o que nos indica que a média e o desvio padrão calculados a partir das duas classes juntas não contêm informação suficiente para nos ajudar a encontrar o limite que melhor as distingue. No entanto, podemos também detectar as relações que se sumariam nos gráficos seguintes⁷:

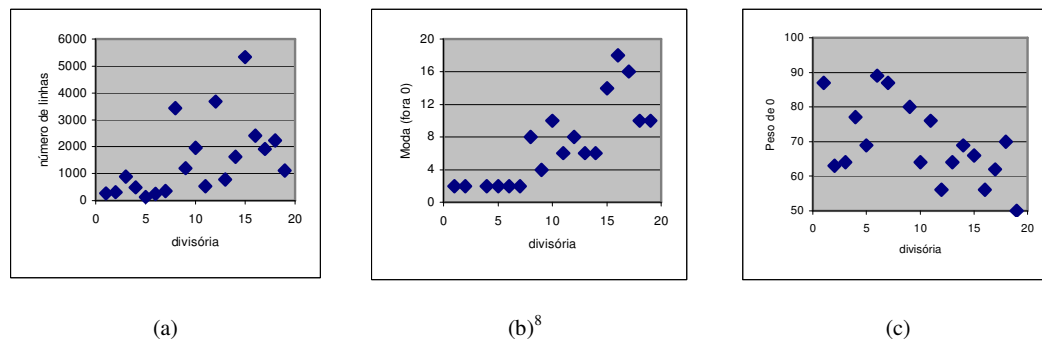


Figura 11. A “divisória ideal” versus o número de linhas (a), a moda excluindo o zero (b) e o peso de zero espaços (c).

Assim, por um lado, relatórios em que o peso das linhas sem espaços interiores é maior (quanto mais próximo de 50%) parecem ter divisórias também maiores; por outro lado,

⁷ Foram analisadas outras medidas de distribuição, como as medidas de assimetria de Bowley ou a *skewness*, bem como a medida de achatamento de Kurtosis (que aqui não tem grande significado, já que a maioria das distribuições é muito assimétrica). Não foram detectadas outras relações marcantes.

⁸ Exclui-se uma observação, que reduziria a legibilidade do gráfico por apresentar valores (4; 62).

relatórios com menos de 500 linhas parecem estar associados a divisórias menores; por fim, os relatórios em que a moda da distribuição (depois de excluídas as linhas sem espaços interiores) é de apenas 2 apresentam divisórias claramente inferiores aos relatórios em que a moda é mais elevada, parecendo notar-se uma correlação positiva entre a divisória ideal e a moda, quando esta é maior do que 2 (que deixa de ser válida quando a moda ocorre na classe de linhas com tabelas). Olhando para os relatórios físicos que aparecem na parte inferior esquerda dos dois primeiros gráficos, notamos que se tratavam daqueles em que o texto está integralmente disposto numa só coluna e que dispõem de menos elementos que não possam ser caracterizados como texto ou tabela (por exemplo, gráficos), sendo também os que são menos prolixos. Estas informações podem vir a ser utilizadas no modelo de decisão que empreguemos.

3.4.1.3. Descritivas das linhas vizinhas

Existe outro ponto muito importante. É que a probabilidade de uma linha pertencer ou não a uma tabela está muito influenciada pela decisão que se tomar em relação às linhas que a envolvem. De facto, uma linha isolada não é uma tabela, ainda que tenha espaços interiores suficientes tal que o pareça; por outro lado, uma linha sem espaços interiores envolvida por linhas com características de tabela será provavelmente pertencente à tabela. Assim, é importante utilizar informação que caracterize as linhas vizinhas na decisão sobre a classificação de dada linha (como também fazem Ng. et al, 1999).

3.4.2. Construção do classificador

A questão que nos ocupa agora é determinar como vamos articular os três níveis de informação necessária à decisão. Um dos factores que mais orientou, neste ponto, foi o facto de constatar-se, no âmbito do nosso contexto específico, que a informação que queremos extrair costuma estar disposta ou em grandes tabelas ou em páginas com várias pequenas tabelas. No gráfico da Figura 12, representada a azul temos, para cada página com tabelas interessantes, a proporção de linhas pertencentes a essas tabelas relativamente ao total das linhas não vazias, representando a cor de rosa a mesma

proporção tomada para as tabelas normais e a cinzento as páginas que têm simultaneamente tabelas interessantes e normais. (O eixo das abcissas está apenas associado ao número de páginas servindo apenas para aumentar a legibilidade do gráfico).

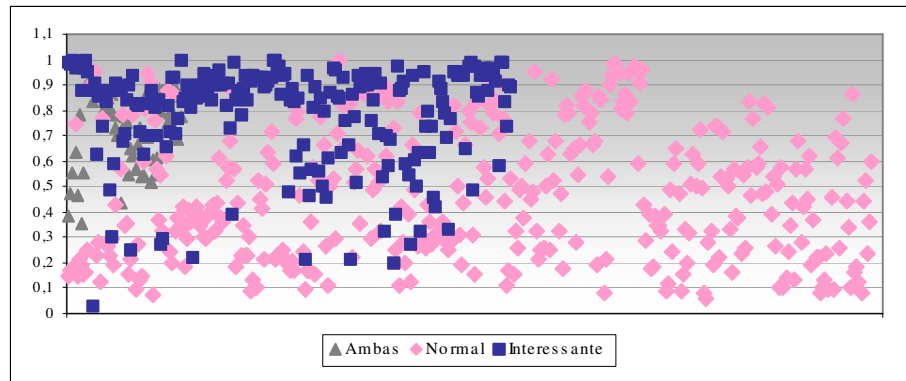


Figura 12. Peso das linhas com tabelas em cada página, por tipo de tabela que a página contém.

Como se tinha previsto, verifica-se que a maior parte das páginas que contêm tabelas interessantes têm elevados pesos de linhas com tabelas (>80%), pesos esses para os quais é relativamente raro encontrarmos linhas com tabelas normais. Com efeito, é comum estas encontrarem-se em páginas mais parcamente preenchidas de tabelas.

Ora, se pudermos encontrar um indicador da concentração de tabelas numa página, podemos utilizá-lo para tornar a nossa análise mais exigente nas folhas que mais nos convêm, conseguindo maior acuidade a distinguir as tabelas que à partida mais nos interessam. Acontece porém que o indicador que analisámos no gráfico acima não pode servir para guiar o sistema, pois está baseado no conhecimento de quais são e aonde se localizam as tabelas que nos interessam, que não é conhecido à partida.

O melhor estimador que pudemos encontrar dessa concentração consiste em identificarmos previamente alguns registos que têm potencialmente características de tabelas tendo em conta o seu número de espaços interiores, contando o número de tais registos por página e comparando com o respectivo número de linhas preenchidas.

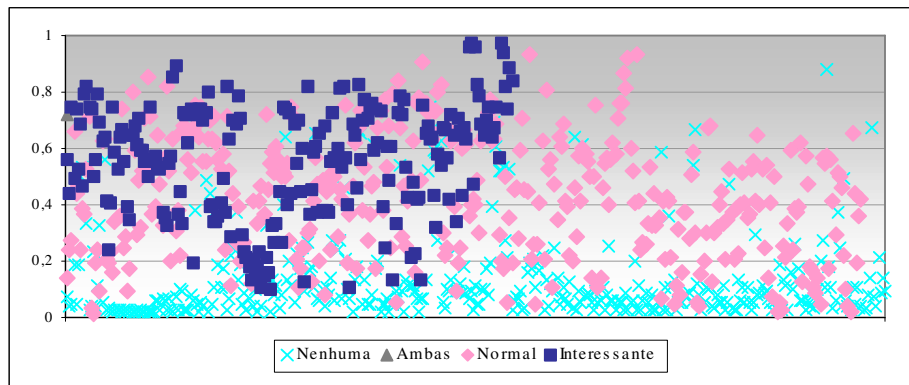


Figura 13. Peso dos candidatos a tabelas em cada página, por tipo de tabelas que a página contém.

Podemos verificar que a maior parte das tabelas interessantes têm valores da estimativa de concentração acima dos 50%, abaixo da qual se situam boa parte das tabelas consideradas normais, que nos é indiferente captar ou não, e das falsas tabelas, sendo que mais de metade destas tem menos de 30% no indicador de concentração.

O uso da concentração por si só não nos traria grande qualidade aos resultados se o método de decisão que aplicássemos fosse igual para todas as páginas. Porém, constata-se que é nas tabelas de maior dimensão que mais facilmente se encontram linhas que não têm características de tabela se vistas isoladamente (como acontece por exemplo na Figura 8, na linha Custos com o Pessoal, que só pode ser detectada como fazendo parte de uma tabela vendo a vizinhança em que está inserida). Em tabelas menores, a maior parte das linhas têm as diversas colunas preenchidas e logo são detectadas pelo sistema com base no número de espaços interiores que contém.

Foi assim que se chegou à conclusão que o método de localização de tabelas deveria operar em duas fases: primeiro, detectando candidatas a partir do número de espaços interiores de cada linha, tendo em conta características da distribuição de espaços brancos por relatório; e depois unindo as candidatas entre si, caracterizando como pertencentes a tabelas as linhas que estejam na vizinhança das candidatas, aceitando uma maior amplitude entre estas nas páginas em que a sua concentração é mais elevada.

A alternativa seria proceder como fizeram Ng et al. (1999), que incorporaram nos descritores de cada linha características de n linhas vizinhas, sendo n fixo. No nosso caso, n varia de página para página, de acordo com o número de pré-candidatas aí contidas, o qual funciona como uma *proxy* da probabilidade de a página conter maioritariamente tabelas. As pré-candidatas são identificados por uma heurística que incorpora as características do documento de origem como um todo.

Assim, em resumo, o nosso método funciona em quatro fases,

1. caracteriza a distribuição de espaços interiores no documento como um todo;
2. compara o número de espaços interiores de cada linha do documento com os valores descritivos daquela distribuição, para identificar as linhas candidatas a tabelas;
3. mede a proporção de candidatas em cada página, para definir o tamanho da vizinhança a que se vai aplicar o classificador;
4. classifica como pertencente a uma tabela cada linha dessa vizinhança, caso ela tenha mais de dois espaços interiores ou tenha um conteúdo suficientemente estreito;

sendo influenciado por três níveis da distribuição de espaços brancos: os do documento como um todo, os da página em que se insere e os da própria linha. O método é robusto à presença de linhas vazias na página, pois define a vizinhança em função do número de linhas preenchidas; é também robusto à presença de linhas dispersas com muitos espaços interiores (como o fim de parágrafos em textos organizados em duas ou mais colunas). Em seguida, apresentamos o algoritmo (a verde os parâmetros manipuláveis):

Nomenclatura:

1. $L \leftarrow$ o conjunto de linhas, l , do documento de origem
2. $P' \leftarrow$ o conjunto de páginas, p' , do documento de origem
3. $Esp_l \leftarrow$ o número de espaços interiores de dada linha
4. $V_l \leftarrow \begin{cases} 1, & \text{se a linha estiver preenchida, i.e., contiver pelo menos um carácter não espaço} \\ 0, & \text{caso contrário} \end{cases}$
5. $Ind_l \leftarrow$ índice, absoluto em relação ao documento, que varia de 1 até $\#\{l \in L : V_l = 1\}$
6. $Md \leftarrow$ a média dos espaços interiores das linhas de $L = \frac{1}{\#L} \sum_{l \in L} E_l$

7. $DP \leftarrow$ o desvio padrão dos espaços interiores das linhas de $L = \sqrt{\frac{1}{\#L} \sum_{l \in L} (E_l - Md)^2}$
8. $C \leftarrow$ o conjunto de linhas que são candidatas a tabelas = $\{c \in L \wedge E_l \geq Md + \gamma DP\}$
9. $P \leftarrow$ o conjunto de páginas, p , com mais de m candidatas = $\{p' \in P' : \#\{c \in p'\} > m\}$
10. $Peso_p \leftarrow$ peso das candidatas no total de linhas preenchidas da página $p =$

$$= \frac{\#\{c \in p\}}{\#\{l \in L : V_l = 1\}}$$
11. $Nova \leftarrow$ sinalizador binário do começo de uma nova tabela
12. $Guar \leftarrow$ a última linha caracterizada como tabela
13. $Ult \leftarrow$ a última candidata analisada
- Algoritmo:**
14. $Nova \leftarrow 1$
15. $Guar \leftarrow 100$
16. $Ult \leftarrow 100$
17. **Para** cada página $p \in P$,
18. **Definir** os seguintes parâmetros em função do $Peso_p$,
19. k $\leftarrow$ distância entre candidatas que se considera aceitável numa tabela
20. i $\leftarrow$ número de linhas preenchidas que, estando antes da primeira candidata detectada de uma tabela, se considera poderem pertencer-lhe
21. f $\leftarrow$ número de linhas preenchidas que, estando depois da última candidata detectada de uma tabela, se considera poderem pertencer-lhe
22. c_0, c_1 e $c_2 \leftarrow$ as 3 primeiras candidatas de p .
23. **Se** $[(Ind\ c_1 - Ind\ c_0 < k \ \underline{\text{e}} \ Ind\ c_2 - Ind\ c_1 < k) \ \underline{\text{ou}} \ (Ind\ c_2 - Ind\ Guar) \leq 1 \ \underline{\text{ou}} \ Peso_p > \underline{y} \ %]$,
 % se as candidatas estiverem suficientemente próximas entre si ou de uma tabela já identificada ou numa página que tem uma grande proporção de candidatas
 então,
24. **Enquanto** $c_2 \in p$,
- % identifica a primeira linha da janela que vamos tratar
23. $Pos \leftarrow$ Primeira linha da página p cujo índice seja $\geq$ a:
$$\begin{cases} Ind\ c_0 - i, \text{ se } (Nova = 1 \text{ ou } Ind\ c_0 - Ind\ Ult > k) \\ Ind\ Ult + f + 1, \text{ caso contrário.} \end{cases}$$
- % trata as linhas da janela definida (entre Pos e $id2+f$)

25. **Para** $l \in p : Ind_{Pos} \leq Ind_l \leq Ind_{c_2} + f$,

26. $Comp_l \leftarrow$ total de caracteres não espaço da linha l

27. $Max(Comp_l) \leftarrow Máximo(Comp_l) : l \in C \wedge l \in p$

% classifica como linhas de tabela as que satisfazem as condições

28. **Se** $Esp_l > z$ ou $Comp_l < x * Max(Comp_l)$, **então** $tabela = sim$.

29. $Guar \leftarrow Pos$

30. $Pos \leftarrow Pos + 1$

31. **Fim de para.**

32. $Ult \leftarrow c_0$

33. $Prox \leftarrow$ candidata seguinte à actual c_2

% define nova janela para tratar na iteração seguinte

34. **Se** $Prox \in p$ e $[Ind_{Prox} - (Ind_{c_2} + f) < k$ ou $Peso_p > \underline{y} \%]$ **então**,

35. $Nova \leftarrow 0$

36. $c_0 \leftarrow c_1$

37. $c_1 \leftarrow c_2$

38. $c_2 \leftarrow Prox$

39. **Se** $Prox \in p$ e $Ind_{Prox} - (Ind_{c_2} + f) \geq k$ e $Peso_p \leq \underline{y} \%$ **então**,

40. $Nova \leftarrow 1$

41. $c_0 \leftarrow Prox$

42. $c_1 \leftarrow$ candidata seguinte à actual c_0

43. $c_2 \leftarrow$ candidata seguinte à actual c_1 .

44. **Se** $Prox \notin p$ **então**,

45. $Nova \leftarrow 1$

46. $p \leftarrow p \in P$ seguinte.

47. **Fim de se.**

48. **Fim de se.**

49. **Fim de se.**

50. **Fim de enquanto.**

51. **Fim de se.**

52. **Fim de para.**

53. **Fim do algoritmo.**

Os parâmetros incorporados no sistema

Existe um parâmetro deste algoritmo que não será definido pelo utilizador e que é necessário decidir que valor tomará. Tal é o γ , apresentado na linha 8. Para tomarmos uma decisão quanto a este valor, decidimos simular diversas alternativas possíveis, medindo para cada uma o erro quadrado médio em que se incorre na estimação da “divisória ideal”. Os resultados apresentam-se na Tabela 13. Com o aspecto de uma parábola com concavidade voltada para cima, o mínimo do erro quadrado médio foi obtido para $\gamma = 0,46$. Optámos ainda por testar os resultados com z igual a 2 e x a $2/3$ (linha 28), para correremos o menor risco possível de criar falsas divisórias a meio de uma dada tabela.

γ	Erro	γ	Erro
1,000	227,15	0,459	135,04
0,667	181,67	0,455	135,07
0,556	144,86	0,450	135,17
0,500	136,70	0,400	139,16
0,470	135,10	0,333	153,07
0,465	135,05	0,250	184,25
0,460	135,03	0,200	210,32

Tabela 13. Simulação do erro quadrado médio para diferentes valores de γ

Os parâmetros manipuláveis pelo utilizador

Quando o objectivo da tarefa de localização de tabelas é a extracção de informação, estamos naturalmente mais interessados em captar com maior grau de certeza as tabelas que contêm a informação alvo. Em contexto genérico, tais tabelas podem ser quaisquer. Porém, no nosso contexto específico, sabemos que as tabelas que procuramos têm maior probabilidade de estar situadas em páginas preenchidas com várias tabelas ou com apenas uma tabela grande.

Os passos 10, 17, 19, 20 e 21 do algoritmo permitem a adaptação do método às características conhecidas das tabelas de que se pretende extrair informação: na décima linha, mede-se o peso das candidatas a tabela no total da página p' em análise, e nas restantes definem-se, para cada página e em função daquele peso, os parâmetros k , i e f com quais o utilizador pode informar o sistema das características das tabelas que pretende. O que ocorrerá é que se definirão maiores vizinhanças nestas páginas, incorrendo-se maior risco de erro encontrar falsos positivos e menores riscos de se

perderem verdadeiros positivos, nas páginas que têm mais indícios de serem de interesse.

Por exemplo:

- caso seja provável que as tabelas interessantes tenham várias linhas contíguas parcialmente preenchidas (ver tabela adjacente), então as candidatas a tabelas (linhas em que estão escritas “A” e “B”) estarão afastadas entre si e deveremos portanto ter elevados valores de k .

A	Existe uma coluna com conteúdos extensos justificados verticalmente, pelo que haverá grandes distâncias entre as linhas que, devido ao número de espaços interiores, são candidatas a tabelas	150
B	também	200

Tabela 14. Exemplo de tabela que implica a opção por um k elevado (mínimo 9).

Temas	Dias de reunião	Meses respectivos
Crescimento demográfico	10	Março
<i>Big bang</i>	15	Abril
Efeito de estufa	20	Maior

Tabela 15. Exemplo de tabela que implica a opção por um i elevado (mínimo 5).

- se for comum os cabeçalhos das colunas das tabelas estarem repartidos em várias linhas, então o i deverá ser tão grande quanto o número de linhas em que se espera que tais cabeçalhos se repartam. No exemplo, as linhas candidatas a tabelas seriam apenas as de dados; para incluir toda a tabela na vizinhança, i deverá ser pelo menos 5.

- se for de esperar que os fins das tabelas interessantes contenham poucas candidatas identificáveis, então f deverá ser elevado.

- se se pensar que a informação de interesse tanto pode estar em tabelas grandes quanto em pequenas, e que estas tabelas pequenas tanto podem partilhar a páginas com outras tabelas quanto com texto simples, então os valores de k , i e f devem ser os mesmos para qualquer *Peso*.

No caso dos relatórios e contas, como se pode observar no gráfico da Figura 13, as tabelas interessantes têm tendência a localizar-se em páginas com concentração acima de 50%; por outro lado, a experiência tem ensinado que a presença de linhas não detectadas como candidatas é menor em tabelas pequenas e que os cabeçalhos

distribuídos por muitas linhas tendem a pertencer a tabelas de maior dimensão. A combinação destes factores levou-nos a testar o método com os seguintes parâmetros.

Se $\text{Peso}_p < 30$ então	Senão se $\text{Peso}_p < 50$ então	Senão se $\text{Peso}_p \geq 50$ então
$k = 1$	$k = 2$	$k = 3$
$i = 1$	$i = 2$	$i = 4$
$f = 1$	$f = 2$	$f = 2$

Tabela 16. Regra de construção dos parâmetros do algoritmo

Por fim, decidimos testar um valor de m igual a 3 (linha 9 do algoritmo). O gráfico da Figura 13 levou-nos a optar por um y igual a 50% (linha 23).

Outros detalhes

Adicionalmente, em lugar de marcarmos cada linha que se creia pertencer a uma tabela apenas com a classificação *sim*, atribuir-lhe-emos um número inteiro n , que terá o seguinte modo de construção: n começa por ser nulo; para cada linha a classificar afirmativamente, verifica se a última linha não vazia anterior foi considerada tabela e está na mesma página; se tal acontecer, então n será o mesmo que o do registo anterior; senão, se a distância entre o registo actual e o último registo identificado como tabela for menor ou igual a um e a página de ambas for a mesma, então n não se altera, senão acrescenta-se-lhe 1.

Isto quer dizer que, se entre duas linhas de tabela houver linhas que não tenham sido detectadas como tabelas, como sejam, uma linha de texto comprida ou um parágrafo de texto com algumas linhas em duas ou mais colunas, o sistema saberá identificar que se trata de duas tabelas distintas.

Esta informação será utilizada apenas como um indicador, pois o princípio de uma tabela será, regra geral mais claramente identificado, pela caracterização das suas primeiras linhas como descritoras de dados. Tem ainda assim algumas vantagens importantes, como poderem-se tomar medidas descritivas de cada uma das áreas de

possíveis tabelas, como sejam, número total de linhas ou média de espaços interiores. Este será um importante *input* para as fases planeadas no capítulo seguinte.

3.4.2.1. Variantes do método de classificação

Método 2

É na linha 28 do algoritmo que verdadeiramente se classifica dado registo como tabela ou não. São classificadas afirmativamente todas as linhas que, pertencendo a uma vizinhança suspeita, tenham espaços interiores ou sejam suficientemente estreitas. A linha 23 do algoritmo impõe uma de três condições para que uma vizinhança, medida a partir de três candidatas a tabelas, c_0 , c_1 e c_2 , possa ser sujeita ao classificador:

- as três linhas estejam suficientemente próximas entre si, [$Ind\ c_1 - Ind\ c_0 < k$ e $Ind\ c_2 - Ind\ c_1 < k$]
- ou a maior das candidatas, $Ind\ c_2$, esteja antes da última linha classificada como tabela, *Guar*, o que pode acontecer quando $Ind\ c_2 + f$ da iteração anterior é maior que $Ind\ c_2$ desta iteração
- ou quando o peso das candidatas a tabelas é suficientemente elevado tal que existem indícios de estarmos perante uma grande tabela ou várias pequenas tabelas, [$Peso_p > y\ \%$].

Esta última condição é menos restritiva que as anteriores, porque se considera que, quando a concentração de candidatas a tabelas é suficientemente alta, existe maior segurança de que as características exigidas na linha 28 são indício de uma tabela. Porém, revisitando a Figura 11b), lembrámos a estável relação que parece existir entre moda da distribuição dos espaços interiores (excluindo os exemplos sem espaços) e a “divisória ideal” sugerida pelo algoritmo J48 de construção de árvores de decisão e mais notámos que para valores suficientemente baixos dessa moda, a primeira divisória sugerida pelo algoritmo era ela também bastante baixa, ou seja, a presença de relativamente poucos espaços já era um bom indicador, nestes casos, de que a linha pertencia a uma tabela.

Decidimos então rever os histogramas da distribuição em cada relatório, apresentados no Anexo 2, e constatámos que há duas grandes localizações possíveis para esta moda, na amostra em causa: em torno do número médio de espaços que, nas linhas só de texto, separam as suas colunas, ou em torno do número médio de espaços nas linhas de tabelas e outros elementos gráficos, e que a localização específica da moda depende do peso relativo de ambas as linhas. Ora, se a moda está muito à esquerda no gráfico, isso pode ser um indicador (e de facto é-o, na nossa amostra) de que o número médio de espaços que, nas linhas só de texto, separam as suas colunas, é zero. Neste contexto, a presença numa linha de ainda que apenas alguns espaços já é, tal como a elevada concentração de linhas candidatas numa tabela, um bom indicador de que estamos perante um elemento gráfico a que há que dar atenção.

Ora, em 6 dos 19 documentos que nos servem de treino, a moda após zero é 2. Que espécie de linha é esta que tem apenas dois espaços interiores e é tão comum? Não há de ser coincidência que esta “moda” aconteça apenas em relatórios pequenos. É que, para que um dado valor se repita muitas vezes em tabelas é necessário que existam muitas linhas de tabelas, pelo menos num contexto em que não haja grande regularidade de comprimentos entre os conceitos que aparecem em cada coluna (como é o caso). Ora, como nós excluímos o zero da distribuição, qualquer pequena parcela de texto que por qualquer motivo tenha dois espaços regularmente por linha, já se torna moda (uma empresa, por exemplo, separa, no índice, o número do capítulo do seu tema com dois espaços; outra costuma colocar dois espaços depois do ponto, em texto corrido).

Baseados neste raciocínio, decidimos alargar a aplicação do passo 23 às candidatas que, estando numa mesma página a mais de k linhas de distância entre si, provenham de relatórios em que a moda é 2⁹. As linhas 34 e 39 têm de ser adaptadas em conformidade.

Método 3

Aqui decidimos, por raciocínio análogo ao anterior, alargar a aplicação do ponto 23 (adaptando as linhas 34 e 39 em conformidade) também aos casos em que a “moda sem

⁹ Devido ao modo como foi construída, a variável só toma valores pares.

zero” seja maior que 50, que já parece ser grande demais para delimitar diferentes colunas de texto, sendo originado pelas linhas que contêm tabelas. Infelizmente, temos uma amostra pequena demais deste tipo de relatório, apenas um, que tem uma frequência invulgarmente alta de tabelas (37% das mais de 2500 linhas). Assim, ainda que nos pareça perfeitamente legítimo, há que ter alguma precaução em termos da para confiança na capacidade de generalização deste ponto do modelo – a verificar perante novos exemplos que obtenhamos no futuro.

Método 4

Decidimos averiguar o que aconteceria se, em vez de considerarmos como pré-candidatas a tabelas as linhas que tenham mais espaços interiores que a média mais 0,46 desvio padrões do documento de origem, linha 8 do algoritmo, utilizássemos um método menos heurístico.

Assim, utilizámos o Método de mínimos quadrados (MMQ) para estimar os coeficientes a aplicar à média (Md) e ao desvio padrão (DP) e ao termo independente, para mais nos aproximarmos da divisória aconselhada pelo algoritmo J48, a partir dos relatórios em que a moda (fora 0) não é 2 nem maior que 50. Obtivemos a seguinte equação: $22.9 + 1.03 \text{ Md} - 0.3 \text{ DP}$. Relativamente a esta, a nossa heurística representa um acréscimo de 18% no erro quadrado e apenas 2% no erro absoluto.

Se porém usássemos o mesmo método tomando como variável independente apenas a moda (fora 0), com base na nossa amostra, obteríamos o modelo $17.9 + 1.4 \text{ Moda}$, que apresenta erro quadrado e absoluto 30% inferiores aos do MMQ anterior (40% menores que o primeiro método testado).

Decidimos testar o que aconteceria se aplicássemos a equação $17.9 + 1.4 \text{ Moda}$ para definição das candidatas, sem esquecer de substituir o valor da variável independente por 2 quando esta for muito grande, por raciocínio idêntico ao já explicado acima. Embora aquela equação tenha sido estimada sem a presença destes valores extremos para a moda, optámos por definir o limite também para estes casos com base nela.

Método 5

Por fim, observámos que, embora em alguns casos o método anterior resultasse em valores mais baixos e portanto mais conservadores para o limite que distingue os candidatos a tabelas, noutros, em poucos, resultava em valores muito superiores, que poderia afectar adversamente a nossa taxa de erro global. É que não há que esquecer que o Método de mínimos quadrados (MMQ) está a minimizar o erro obtido em 12 observações em que todas têm o mesmo peso, porém, na nossa amostra, elas afectarão o erro global de modo proporcional ao número de linhas de cada relatório (e no Universo, também ao peso de cada tipo de relatório no total). Assim, decidimos adaptar a linha 8 do algoritmo para avaliar o que aconteceria se utilizássemos como delimitador o mínimo das duas opções possíveis: $\text{Mínimo} \{ 17.9 + 1.4 \text{ Moda} ; \text{Md} + 0,46 \text{ Dp} \}$.

3.4.3. Metodologia de avaliação dos resultados

3.4.3.1. No contexto específico – captar as tabelas interessantes

Medidas de avaliação a usar

A nossa experiência no caso de extracção de informação é algo peculiar de avaliar. Trata-se de um problema de classificação em que estamos particularmente interessados em encontrar todos os elementos de um subconjunto de uma das classes, o das tabelas interessantes. No entanto, a informação com que se pode caracterizar uma linha antes de esmiuçar o seu conteúdo não nos permite distinguir totalmente, por agora, as tabelas interessantes das normais.

Na verdade, como foi dito no início desta secção 3.4., o principal objectivo desta fase não é termos muita certeza que o que classificamos como tabela de facto o é, é o termos o maior grau de certeza possível de que as linhas que classificamos como não tabelas não são tabelas interessantes. Qualquer diminuição na dimensão do nosso documento é porém bem vinda, pois os processos que se seguem a esta primeira localização,

nomeadamente o de segmentação de cada linha em colunas, serão capazes de aumentar em muito a dimensão do nosso problema.

Do que ficou dito, ressalta que há duas grandes medidas para avaliar o nosso sucesso:

-  a recuperação, isto é, a sensibilidade do sistema para captar o pretendido (dada pelo número de classificações correctas da classe objectivo, ou verdadeiros positivos, sobre o número total de elementos dessa classe, i.e., verdadeiros positivos mais falsos negativos) – este objectivo é para nós imprescindível;
-  a economia computacional gerada, medida pelo peso dos registos que deixarão de entrar na engrenagem de decisão – este objectivo é para nós desejável, mas estamos dispostos a abdicar de pouca recuperação para termos mais economia.

Note-se que a economia computacional será maior nos documentos em que o peso de linhas com tabelas é menor, sendo incorrecto utilizar esta medida para comparar o desempenho dos métodos em diferentes amostras, mas não para comparar diferentes métodos entre si. Por sua vez, não faz muito sentido avaliar os métodos pela precisão, pois embora se pretendam as tabelas interessantes, o sistema não está ensinado a evitar as tabelas normais e o total dos detectados reflectirá isso mesmo.

Por outro lado, apesar de o sistema só classificar duas classes, estamos de facto perante três, devido às diferenças de custo do erro a elas associado. Assim, para calcularmos aquelas medidas, teremos de adoptar critérios pouco habituais para definição dos falsos positivos (FP), verdadeiros positivos (TP), falsos negativos (FN) e verdadeiros negativos (TN), que se sumariam na tabela a seguir:

Previsto	Real		
	Não é tabela	É tabela	
		Normal	Interessante
Não é tabela	TN	TN	FN
É tabela	FP	FP	TP

Tabela 17. Cálculo dos elementos necessários à determinação das medidas de avaliação.

sendo, $\text{Recuperação} = TP / (TP+FP)$ e $\text{Economia} = (TP + FP) / (TN + FN + FP + TP)$. Assim, a recuperação é medida em relação apenas às tabelas interessantes.

Custos associados aos diferentes tipos de erro

Como foi dito no início da secção 3.4., temos custos diferentes associados aos diferentes tipos de erros. Com efeito, é-nos indiferente captar uma linha de tabela normal: se por um lado, é algo para o que preparamos o sistema – captar tabelas, por outro trata-se de informação que não necessitamos e que mais cedo ou mais tarde será descartada.

Além disso, o custo de o método captar uma determinada linha como sendo tabela sem esta o ser implica que se manterá no processo um registo desnecessário, que passará por todos os procedimentos mais dispendiosos seguintes. Ainda assim, isso é menos custoso que a não detecção de uma linha de tabela, que só pode ser corrigida se se puder vir a detectar se dada tabela interessante foi completamente extraída e, caso não tenha sido, resgatar do conjunto das linhas rejeitadas a envolvente da tabela em causa.

É porém difícil medir *a priori* estes custos relativos, pois estão relacionados com factores que não podemos medir, até porque não são iguais para todas as linhas:

-  se o erro ocorrer numa tabela de que há que extrair informação, o seu custo é maior que numa tabela que não contém informação interessante;
-  diferentes linhas de uma tabela têm diferentes importâncias para a capacidade de delas se extrair informação;
-  a probabilidade de o sistema ser capaz de, *a posteriori*, detectar se captou todas as linhas relevantes da tabela, será à partida maior quando as linhas em falta se situem no meio das tabelas e contenham valores numéricos relacionados entre si;
-  a eficácia com que o sistema é capaz de resgatar as linhas em falta do conjunto das anteriormente rejeitadas – factor indeterminável nesta fase do trabalho.

Por isso diremos apenas que o custo de não se encontrar uma linha de uma tabela é $x > 1$, tal como explicitado na tabela abaixo. Avaliaremos qualitativamente, de acordo com aqueles estes quatro factores, a gravidade da não detecção de cada linha de tabela e identificaremos a causa do problema. Depois, averiguaremos quantos FP teríamos de aceitar para ver o problema corrigido, decidindo, perante isto, se compensa ou não adaptar o método.

Previsto	Real		
	Não é tabela	É tabela	
		Normal	Interessante
Não é tabela	0	0	$x > 1$
É tabela	1	0	0

Tabela 18. Custos associados aos erros de classificação dos registos.

3.4.3.2. No contexto genérico – captar qualquer tabela

Se o que estamos aqui a propor é um classificador que determine se dada linha de um documento pertence ou não a uma tabela e dizemos, como dissemos no ponto 3.3, que “queremos que o método seja o mais geral possível”, então temos de avaliar o método em termos da sua capacidade de captar toda e qualquer tabela, eliminando o maior número possível de linhas que não são tabelas.

Continuaremos a utilizar as mesmas duas medidas de qualidade dos resultados, desta vez contando os TN / FN / FP / TP de acordo com a tabela abaixo.

Previsto	Real	
	Não é tabela	É tabela
Não é tabela	TN	FN
É tabela	FP	TP

Tabela 19. Cálculo dos elementos necessários à determinação das medidas de avaliação.

As grandes questões que queremos responder, no âmbito da avaliação do método em contexto genérico são: Que economia máxima é possível para uma recuperação o mais próximo possível de 100%? Que parâmetros hão-de utilizar-se para obter isso?

3.4.3.3. Conjunto de teste

Para averiguar a qualidade da classificação, aplicaremos o algoritmo aos 19 relatórios de que dispomos, os nossos exemplos de treino. Porém, queria lembrar que o método proposto não se baseia em modelos formais, que são o resultado directo dos exemplos sobre que trabalham, sobre os quais não se conhece a representatividade em relação ao Universo. O método proposto é sim o resultado da aplicação de um entrançado de heurísticas, baseadas num conhecimento empírico da apresentação da informação neste tipo de documentos, nas suas excepções, detectadas não só neste conjunto, mas também numa amostra mais ampla de relatórios em papel, e na constatação, intuitivamente bastante generalizável (e estatisticamente consubstanciada no Apêndice 1) que, regra geral, a distribuição de espaços brancos é diferente entre tabelas e texto.

Ainda assim, é inevitável que as decisões tomadas quanto aos parâmetros tivessem por base os dados disponíveis e as excepções encontradas, pelo que o teste neste conjunto não é tão fiável quanto se fosse num conjunto nunca antes visto. Por isso, depois de elegermos os melhores métodos com base nos resultados da primeira fase de testes, aplicá-los-emos a um conjunto fresco de elementos de teste:

Futebol Clube do Porto – Futebol SAD	2000 / 2001
Sonae Indústria SGPS, SA	2000
Sonae.com	2000

3.4.4. Resultados experimentais no conjunto de treino

3.4.4.1. Avaliação em contexto específico

No Anexo 3 apresenta-se um quadro que sumaria os resultados da aplicação dos diversos métodos aos exemplos disponíveis. Os cinco métodos revelaram-se bastante parecidos em ambas as medidas, com a economia média a variar entre 75% e 77% e a sensibilidade média mantendo-se entre 89% e 91%, para o total das 87.843 linhas dos 19 relatórios.

A escolha do melhor dos métodos básicos

Evidentemente, existe uma relação de compromisso entre as medidas de avaliação, que é necessário equilibrar. Na figura seguinte, representamos os diferentes métodos num gráfico em que os eixos captam as nossas medidas alvo.

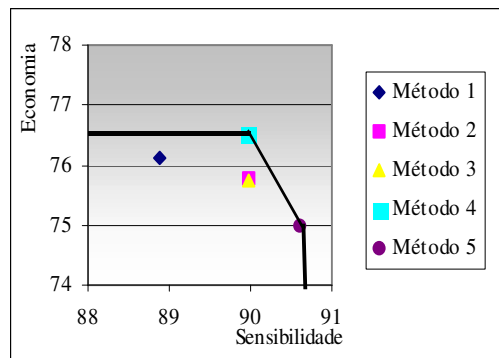


Figura 14. Escolha do método - fronteira de eficiência.

Como se pode verificar, são os métodos 4 e 5 que definem a nossa fronteira de eficiência. Com efeito, qualquer dos outros implicaria abdicar de um dos objectivos sem ganhar mais no outro. Dado que para nós sensibilidade é mais importante que economia, parece sensato abdicar de 1,62 p.p. de economia para ganhar 0,67 p.p. em sensibilidade. Por

outro lado, o método 5 consegue detectar todos os registos interessantes de 8 relatórios (que representam 45% do total das linhas e incluem o maior e o mais pequeno documento da nossa amostra), contra 6 no método 4. Assim, parece-nos que o método 5 é o melhor dos métodos apresentados até aqui.

Análise crítica do funcionamento do melhor método

O gráfico abaixo mostra como se comporta o método 5 nos diferentes relatórios, apresentando a vermelho o desempenho obtido no total dos relatórios.

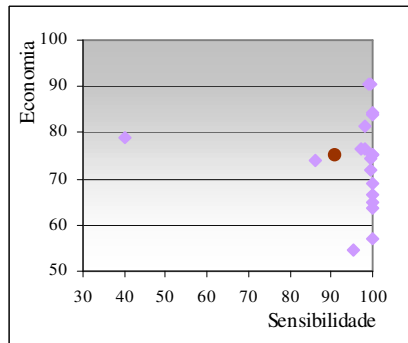


Figura 15. Desempenho do método escolhido por relatório.

Verifica-se que em 17 dos 19 relatórios a sensibilidade foi acima dos 95% (em 8 igual a 100%) e em 11 temos economias acima dos 75%. Não existe nenhuma tabela interessante que tenha ficado inteiramente por detectar. Estes resultados parecem-nos muito satisfatórios. Mas poderão ser melhorados?

Para responder a esta questão, decidimos analisar os erros cometidos pelo método nos diferentes documentos e averiguar as possíveis soluções. Note-se que

Tipo de erro Peça contabilística	a	b	c	d	e	Total
	Cabeçalhos	Totais	Linhas estreitas	Linha sem espaços	Abarcadoras	
Passivo e capital próprio		4				4
Movimentos em provisões	5					5
Demonstração de resultados financeiros		1	6			7
Empresas relacionadas	14				868	882
Demonstração de resultados por funções			10			10
Movimentos no capital próprio				1		1
Total	19	5	16	1	868	909
Número de documentos em que o erro ocorre	4	5	2	1	2	8
Solução	Aumentar <i>i</i>	Aumentar <i>f</i>	--	Pós processamento	Aumentar <i>k</i> e <i>f</i>	

Tabela 20. Tipologia dos erros cometidos pelo método 5.



Erro a

O problema verificou-se num conjunto de tabelas que têm cabeçalhos algo amplos e concomitantemente cujas primeiras linhas de dados não estão preenchidas. A solução possível para este erro seria aumentar o parâmetro i em 4 unidades quando o peso das candidatas a tabelas na página for menor que 30% e em 5 quando for menor que 50%. Evidentemente, esta seria uma adaptação que resolveria os problemas ocasionados por este erro neste subconjunto específico. O mais coerente nesse sentido seria aumentar em 5 o parâmetro i , qualquer que seja a página. Isto tornaria o método mais generalizável.

Para ter noção dos custos associados a esta decisão, decidimos ver o que aconteceria se aumentássemos i em apenas uma unidade, apresentando os resultados na barra vermelha do Anexo 3. Podemos constatar que o custo de cada novo verdadeiro positivo captado por esta via é 6 falsos positivos. Podemos ainda constatar que 90% do benefício dá-se nos documentos da EDP, cuja causa é o erro d. Nos restantes documentos, o custo por cada novo verdadeiro captado é de 42 falsos. Isto é um custo mesmo muito elevado, tanto mais que corresponde a aumentar i em apenas uma unidade, e não 5.

Assim, pretendemos averiguar no futuro o custo de corrigir este problema a posteriori, apenas para o conjunto das tabelas que haja motivos para crer conterem a informação alvo de extracção. A detecção de um título na zona acima da tabela poderia contribuir para a confirmação de se se encontraram todos os cabeçalhos ou, caso contrário, permitir o resgate das linhas entre a tabela e o título para aperfeiçoamento da delimitação das poucas tabelas afectadas por este problema. Esta solução parece mais coerente.



Erro b

Num total de cinco documentos e quatro tabelas, não foram detectadas as linhas finais. Isto acontece porque estas linhas têm denominações muito extensas, tal que não contêm suficientes espaços interiores para serem consideradas candidatas e situam-se a mais de f linhas da última candidata seleccionada. A alteração nos parâmetros que resolveria

este problema seria aumentar em uma unidade o f , o que permitiria resgatar 4 dos 5 erros detectados. O custo associado a esta decisão é de 14 falsos por cada verdadeiro, mas 90% dos verdadeiros registam-se nos documentos da EDP e apenas 3 dos 5 mencionados na tabela 19 seriam resolvidos.

Existe porém uma alternativa mais económica. É muito comum as tabelas que apresentam números descreverem na última linha um conceito começado por “Total”. Isto é verdade tanto para as tabelas de cariz estritamente contabilístico, como também para outros contextos. Se por um lado, esta solução é menos genérica, pois não soluciona o problema das tabelas não numéricas, por outro ela é mais generalizável, pois permite detectar o fim da tabela mesmo quando um aumento de 1 em f não é suficiente para resolver o problema.

Decidimos averiguar os custos associados com identificar todas as linhas que contenham a expressão “Total”, ou no início da linha ou precedida de pelo menos dois espaços. São identificados 328 exemplos no total das 87.843 linhas. Destes, 165 pertencem a tabelas interessantes, dos quais 5 não foram detectados pelo método 5 (ou seja, a totalidade dos erros originados por esta causa); 138 pertencem a tabelas normais, das quais 10 não foram detectados pelo método 5; e 25 não pertencem a tabelas, dos quais 17 não foram detectadas como tabelas. Assim, o custo de cada verdadeiro incorporado por esta via seria igual a $27 / 5 = 5,4$, o que é bastante mais favorável que a alternativa anterior. Claro que, de entre o conjunto de linhas que contêm a expressão “Total”, apenas seriam passadas à fase seguinte as linhas que estejam na mesma página de tabelas detectadas. Caso se verifique que alguma destas linhas não foram detectadas automaticamente pelo método 5 e são de facto de uma tabela, poder-se-á facilmente resgatar as linhas que medeiam entre ela e a última detectada.

Poderia ser interessante alargar o conjunto de palavras chaves, como por exemplo com “resultado”, uma vez que esta é uma expressão alternativa a “total” e amplamente utilizada nas tabelas contabilísticas para marcar o seu meio ou fim.



Erro c

Existe um conjunto de 16 linhas de 2 documentos que não são detectadas por serem estreitas. Estas linhas provêm de duas tabelas distintas, cada uma de um relatório da EDP (1997 e 2000), porém outras linhas da mesma tabela são identificadas, o que abre caminho para a possibilidade de se poderem mais tarde resgatar estas linhas (quando se notar que uma tabela sem linha de cabeçalhos se encontra abaixo e suficientemente próximo de uma tabela que os tem, dever-se-á analisar o espaço que as medeia para ver se serão uma só tabela). Infelizmente, o nosso método não está, nesta fase, capacitado para tratar tabelas que, sendo amplas, contêm conjuntos contíguos de linhas estreitas. Em relação a tabelas estreitas, o nosso método apenas está capacitado para encontrar as que estejam alinhadas horizontalmente com outros elementos gráficos.



Erro d

Este erro ocorre em apenas uma linha interessante, que é bastante comprida (tem mais que $2/3$ do comprimento total da maior candidata detectada na sua vizinhança de referência) e não tem nenhum par de espaços contíguos (nem nenhum dado), logo é rejeitado pelo passo 27. do algoritmo. É bastante simples corrigir este tipo de erro à posteriori, pois o resultado no sistema é existirem duas tabelas contíguas espaçadas de uma linha preenchida, em que uma delas tem pelo menos uma linha de referência e a outra não. Falaremos deste passo de pós-processamento no capítulo seguinte.



Erro e

Existem dois documentos em que o desempenho é bastante mau, apresentando sensibilidades de 40% e 86% respectivamente. Estes resultados acontecem em dois relatórios da mesma empresa em dois anos consecutivos e localizam-se numa tabela específica - a descrição das empresas relacionadas. Um extracto das tabelas problemáticas é apresentado na Figura 16.

valores dos parâmetros. (O documento contém 9.807 linhas, das quais 1.271 pertencem a tabelas interessantes, linhas de tipo 2, e 1.303 a tabelas normais, linhas tipo 1).

	Linha	k por método	$k = 5$	$k = 10$	$k = 20$	$k = 30$	$k = 40$
$m = 3$	0	362	465	693	908	943	943
	1	1176	1242	1260	1276	1276	1276
	2	510	690	843	1078	1106	1106
Sensibilidade		40	54	66	85	87	87
Economia		79	76	71	67	66	66
$m = 2$	0	529	814	1011	1150	1150	1150
	1	1223	1253	1271	1280	1280	1280
	2	925	902	927	1106	1106	1106
Sensibilidade		73	71	73	85	87	87
Economia		73	70	67	64	64	64

Tabela 21. Manipulação dos parâmetros do algoritmo para correcção dos problemas da EDP2000.
 Legenda: 0. linhas que não pertencem a tabelas; 1. linhas de tabela não interessantes; 2. linhas de tabelas interessantes.

Os melhores resultados que se obtêm neste documento são conseguidos com $k = 30$ e $m = 3$. Relativamente ao método 5, eles implicam, por cada verdadeiro positivo adicional, que se aceite 1,14 falsos positivos, o que parece uma troca aceitável.

Decidimos então manipular os valores de i e f . Para $f = 31$ e $i = 5$, são seleccionadas 3.838 linhas, das quais 1.280 não são tabelas, 1.288 são tabelas normais e 1.270 são tabelas interessantes. Isto representa, relativamente ao método com $k = 30$ e $m = 3$, captar 1,82 falsos positivos adicionais para obter mais 1 verdadeiro positivo, o que ainda parece aceitável. Com estes valores, fica por captar 1 linha de tabela interessante que não contém dados e que se situa no meio da tabela, erro d, (logo que é fácil resgatar a posteriori); e 15 linhas de tabelas normais estreitas, que estão em páginas em que foram detectadas menos de 3 candidatas.

C. se, após a interpretação, o sistema detectasse que aquelas peças contabilísticas não foram de todo encontradas, poderia:

1. alertar o utilizador para que este indique manualmente as linhas / páginas em que se encontra a informação de interesse;

2. ou pedir ao utilizador que confirme a presença no documento da informação em falta e experimentar parâmetros progressivamente mais permissivos até encontrar a informação em falta entre a anteriormente descartada;
3. ou seleccionar de uma vez todas as páginas que continham alguma candidata rejeitada e procurar aí as porções da *template* que não foi possível preencher na primeira passagem, precedendo eventualmente este procedimento de uma confirmação pelo utilizador de que a tabela existe no documento.

Temos de decidir se queremos adaptar os nossos parâmetros para os valores mais permissivos da alternativa B. para que possam lidar com esta situação em qualquer relatório, sacrificando economia para aumentar sensibilidade; ou se preferimos que ou o utilizador ou o próprio sistema detecte a inadequação dos parâmetros para lidar com tabelas interessantes específicas de dado relatório, procedimento mais próximo da alternativa C.. Dado que é muito invulgar, nas tabelas de relatórios e contas, a opção por uma apresentação deste tipo – a inserção de uma longa descrição de actividade, e dado que genericamente estamos interessados em captar números ou nomes de empresas, que raramente ocupam mais de duas linhas, parece-nos impróprio adaptar o modo de tratar todo e qualquer relatório a esta especificidade da EDP. Isto é tanto mais verdade quanto as linhas que causam a perturbação serem absolutamente irrelevantes em termos da informação que se quer extrair.

Mas para ajuizarmos mais cuidadosamente do custo / benefício associado à opção por valores mais permissivos dos parâmetros do modelo, aplicámos o método assim parametrizado à totalidade dos documentos do conjunto de treino. A economia computacional que atingiríamos seria de 62,6% para uma sensibilidade de 99,9% - em 18 dos 19 relatórios, a sensibilidade é acima dos 99%. Assim, teríamos de estar dispostos a tratar 10,51 falsos positivos a mais para captarmos um verdadeiro positivo adicional. Evidentemente, 97% das vantagens deste método ocorrem nos relatórios recentes da EDP: se medirmos o custo apenas nos restantes documentos, ele passa a 36,4 por verdadeiro adicional. Este parece-nos um custo bastante elevado, tanto mais quando existem alternativas: haverá, de futuro, que averiguar o desempenho da

alternativa C. e reavaliar a questão, mas de momento parece-nos que o método 5 continua a ser preferível.

Outros detalhes

Existe um conjunto diminuto de linhas que são agrupadas em pseudo-tabelas de duas ou menos linhas. Isto acontece, por exemplo, quando uma linha que não é tabela pertence a uma vizinhança válida para o algoritmo, pelo facto de estar suficientemente próxima de uma tabela, e é classificada como “com características de tabela” pelo facto de ser suficientemente estreita ou conter espaços interiores (ponto 27 do algoritmo). Porém, entre esta linha e o início real da tabela existem outras linhas que não cumprem com os requisitos do ponto 27, sendo rejeitadas enquanto tabelas. Assim, quando o primeiro registo da tabela real é classificado afirmativamente, é considerado como uma tabela diferente. Poderiam assim ser reclassificados como não pertencendo a tabelas 99 registos, dos quais 28 são títulos e 6 são notas de rodapé. Esta última propriedade será revisitada no capítulo seguinte.

3.4.4.2. Avaliação em contexto genérico

A opção pelo método 5 feita no ponto anterior é válida no nosso contexto específico, em que conhecemos as características tipo das tabelas que procuramos. Porém, se queremos um método facilmente adaptável a qualquer contexto, precisamos saber quais os parâmetros que podemos utilizar para captar toda e qualquer tabela. Interessantemente, o contexto contabilístico é um bom ambiente para testar esta capacidade, pois os relatórios e contas utilizam uma grande diversidade de elementos gráficos, como foi algumas vezes referido no capítulo 2.

No Anexo 4 apresentamos os valores obtidos para as medidas de desempenho escolhidas (no ponto 3.4.3.2.) para os diferentes relatórios e métodos testados. Resumimos isso mesmo no quadro seguinte.

Parâmetros	Método 5	Método 5 a)	Método 5 b)	Método 9	Método base
m	3	3	3	3	0
k	varia	varia	varia	31	-
i	varia	varia + 1	varia + 1	5	-
f	varia	varia	varia + 1	31	-
Resultados					
Sensibilidade	90,1%	90,4%	91,2%	98,5%	99,9%
Economia	75,0%	74,7%	73,9%	62,6%	47,7%
Documentos aonde sensibilidade $\geq 95\%$	12/19	12/19	13/19	17/19	19/19

Tabela 22. Sensibilidade e economia para diversos valores dos parâmetros manipuláveis pelo utilizador.

O método base consiste em escolher todos os registos das páginas em que exista pelo menos uma candidata. É ainda assim interessante verificar quais as tabelas que o método base não é capaz de captar por manipulação dos valores de m , k , i ou f , pois isso avalia o método no seu âmbito – na capacidade de detectar candidatas – e elucida sobre os aspectos do método que têm de ser melhorados de futuro. No total, o método deixa passar 9 registos com tabelas normais provenientes de duas tabelas. Ambas são bastante estreitas e estão em páginas de resto preenchida com texto.

Com o método base, conseguimos, relativamente ao método que utiliza os valores mais permissivos dos parâmetros, um acréscimo de 1,4 p.p. em sensibilidade com um custo de 14,9 p.p. de economia, i.e., para conseguirmos um verdadeiro adicional, temos de estar dispostos a aceitar 53 falsos: o custo é muito elevado. O método 9 é portanto eleito o melhor método: relativamente ao método 5, ele representa uma melhoria 8,1 p.p. em sensibilidade contra 12,1 p.p. em economia, o que é um custo aceitável. No total, perdem-se 251 registos. O gráfico seguinte representa as respectivas causas.

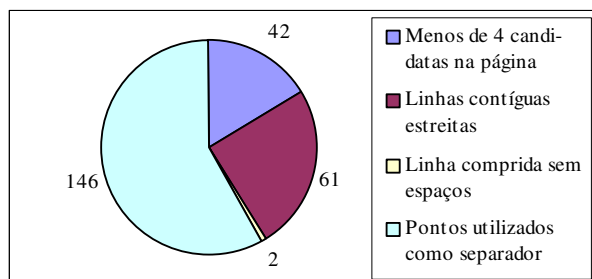


Figura 17. Tipologia de erros do método base

3.4.5. Resultados experimentais no conjunto de teste

Aplicaremos aos três exemplos de teste de que dispomos os métodos que apresentaram melhores resultados entre os de contexto específico (que tem por objectivo rejeitar linhas captando o maior número possível de linhas de tabelas interessantes) e os de contexto genérico (que pretendem rejeitar linhas captando o maior número possível de linhas de qualquer tabela). Os resultados podem ser vistos na tabela seguinte:

	FCP 2001	Sonae indústria 2000	Sonae.com 2000
Características dos documentos			
Total de linhas sem tabelas	2.324	2.742	2.322
Total de linhas de tabelas normais	245	150	301
Total de linhas de tabelas interessantes	264	700	422
Em contexto específico – Método 5			
Não tabelas captadas	99	198	185
Tabelas normais captadas	233	121	290
Interessantes captadas	262	694	422
Economia	79,0%	71,8%	70,5%
Sensibilidade	99,2%	99,1%	100,0%
Peça contabilística com linhas não detectadas	Lista de empresas relacionadas	Demonstração de resultados por funções	
Causa provável da não detecção	Cabeçalhos acima de $c_0 - i$	Últimas linhas a mais de $c_2 + f$	
Solução possível	Aumentar i em 2 ou pós-processamento	Aumentar f ou <i>query</i> “Total”	
Em contexto genérico – Método 9			
Não tabelas captadas	432	678	265
Tabelas normais captadas	245	146	298
Interessantes captadas	264	700	422
Economia	66,8%	57,6%	67,7%
Sensibilidade	100,0%	99,1%	99,6%
Causa provável da não detecção		Em página com apenas 3 candidatas	Tabela com apenas 2 linhas de dados
Solução possível		Diminuir m em 1	Diminuir m em 1

Tabela 23. Avaliação do método em exemplos independentes.

Os resultados são muito bons em termos de sensibilidade em ambos os métodos (tanto mais se considerarmos que a junção dos elementos seleccionados pela *query* que capta as palavras normalmente presentes no fim de tabelas em contexto contabilístico permite elevar a sensibilidade a 100% e manter uma economia de 71,3%, no relatório da SONAE Indústria 2000). A perda de linhas interessantes, num e noutro método, ocorre apenas nos casos já previamente identificados e o método incorporará mecanismos para lidar com estas situações em fases posteriores da análise.

3.4.6. Trabalho futuro

Olhando para a última coluna da Tabela 20, que apresenta o número total de linhas não captadas por peça contabilística, podemos notar certas regularidades, que ou exigem trabalho futuro, ou exigem que o sistema incorpore mecanismos para garantir uma recuperação simples das mesmas. O sistema terá mecanismos para tratar, por exemplo, o problema que afecta a demonstração de resultados por funções (da EDP 2000), que discutiremos no ponto 4.8.1.. Foquemo-nos pois nos erros que exigem trabalho futuro.

Neste caso estão as tabelas com empresas do grupo, incluídas ou excluídas da consolidação, em que se utiliza um grande número de formatos. Apesar de a manipulação dos parâmetros manipuláveis pelo utilizador permitir resolver o problema no conjunto dos exemplos de treino, gostaríamos de melhorar o desempenho no tratamento destas tabelas. Assim, uma pista que poderemos seguir é capacitar o modelo para perceber a possibilidade de estar diante de uma tabela deste tipo e entender isso como um factor de risco, adaptando os seus parâmetros para valores menos exigentes nas páginas em causa (em vez de no documento como um todo). Para tal, bastaria possivelmente ter em conta o número de nomes de organizações localizados por página – acima de determinado nível seria um factor de risco. Wakao (1996) menciona um algoritmo que consegue detectar a presença de nomes de organizações com desempenhos que parecem bastante bons, tendo a sexta Message Understanding Conference (MUC-6) lançado um desafio sobre este tema.

À medida que a amostra for crescendo no tempo, seria interessante investigar a capacidade de definição do limite para identificação das candidatas a tabelas a partir do valor ideal para o ano anterior ou eventualmente de outras do mesmo grupo económico. Depois de extrair toda a informação de um relatório, o sistema poderia determinar qual teria sido o limite ideal e utilizar este valor na estimação dos exercícios seguintes. Isto corresponde a acrescentar uma capacidade de aprendizagem ao método. Haveria que medir os custos e benefícios associados em termos de desempenho.

Por fim, o método não está preparado para tratar tabelas que usem caracteres que não espaço no intervalo entre o fim da coluna de referência e o início da coluna de dados:

	2000	1999
Despesas de instalação	1.000.000	500.000
Despesas de investigação e desenvolvimento	1.500.000	700.000

Tabela 24. Exemplo de tabela que o método não está programado para captar.

A adaptação do método a estes casos não é complexa: bastaria identificar os símbolos que costumam ser utilizados como separadores de coluna e contar o número de ocorrências contíguas de tais símbolos por linha, da mesma forma que se contam os espaços. A generalidade das linhas terá certamente valor zero, pelo que as linhas em que existam valores ainda que ligeiramente elevados dessa variável serão de classificar como “tem características de tabela”, não sem antes substituir ditos símbolos por igual número de espaços, para permitir o mesmo tratamento nos passos seguintes do método.

3.4.7. Conclusão

Vimos como foi criado um método que, tendo por *input* as linhas de um texto contendo tabelas, as classifica em duas classes: “tem características de tabela ou não”. São utilizados, na decisão de classificação de cada linha, aspectos da distribuição de espaços contíguos interiores das linhas medidos a quatro grandes níveis: 1. na própria linha; 2. no documento de origem; 3. na respectiva página; 4. e na sua vizinhança estrita.

O método está afinado para detectar tabelas grandes ou que, sendo pequenas, partilhem a página com outras tabelas pequenas, conseguindo, no nosso contexto específico e utilizando os parâmetros *default* que serão incorporados no sistema, em 17 dos 19 documentos, descartar mais de 74,4% das linhas mantendo 99,6% das linhas das tabelas com informação passível de extracção. Nos restantes dois relatórios, por manipulação dos parâmetros *default* é possível descartar 62,3% dos registos, conservando 100% dos desejados. No conjunto de teste, em contexto específico, a economia média foi de 73,6%, que corresponde a uma sensibilidade de 99,4%; em contexto genérico, alcançou-se uma economia de 63,6%, com sensibilidade de 99,7%.

Por fim, dado que na classe das tabelas estão misturadas linhas que nos é indiferente conseguir captar ou não - aquelas que não contêm a informação que se pretende extrair, e porque o objectivo final da tese é a extracção de informação, fizemos uma pequena adaptação ao modo de cálculo dos valores necessários para determinar as medidas de avaliação dos procedimentos: sensibilidade, precisão e economia computacional.

A tabela a seguir tem por objectivo facilitar a localização dos métodos propostos nesta tese no conjunto de métodos descritos no capítulo 2, fazendo a ponte entre os dois capítulos (os pontos na secção classificador significam que aqueles métodos foram usados como ferramentas, mas não são centrais no esquema proposto).

		Rus e Summers	Douglas et al.	Tupaj et al.	Shamillan	Pyreddi e Croft	Ferguson	SAIC	Kornfeld e Watteamps	Kieninger	Hu et al	Ng et al.	Chen et al.	
		1994	1995	1996	1997	1997	1997	1997	1998	1998	1999	1999	2000	2002
Informação tida em conta	Distribuição de espaços	x	x		x	x	x	x				x...		x
	Posição relativa das palavras			x						x				
	Medidas de semelhança		x		x						x		x	
	Palavras chave			x			x							
	Análise lexical	x											x	

		Rus e Summers	Douglas et al.	Tupaj et al.	Shamillian	Pyreddi e Croft	Ferguson	SAIC	Kornfeld e Wattercamps	Kieninger	Hu et al.	Ng et al.	Chen et al.	
		1994	1995	1996	1997	1997	1997	1997	1998	1998	1999	1999	2000	2002
Método	Busca de delimitadores	x	x	x								x		x
	Agrupamento de blocos									x	x			
Classificador	Heurísticas		x	x		x		x					x	x
	Padrões			x						x				
	Base estatística	x												.
	Maximização de dada medida				x		x				x			
	Modelo formal											x		.
Uso de parâmetros	Pré definidos		x	x		x		x					x	x
	Definidos por utilizador	x				x								x
Aplicabilidade do método	Genérica	x	x	x		x		?		x	x	x		x
	Específica				x		XML		XML				HT ML	

Tabela 25. Sumário das estratégias de localização de tabelas.

4. Planeamento do funcionamento das restantes tarefas

4.1. Definição do modelo físico das tabelas, parte I - segmentação em colunas

4.1.1. Segmentação das linhas em células

4.1.2. Repartição das células pelas suas colunas

4.1.3. Validação do número total de colunas por tabela

4.2. Complementos à localização - eliminação das falsas tabelas que restam

4.3. Análise funcional

4.4. Definição do modelo físico das tabelas, parte II – células abarcadoras

4.4.1. Identificar células que abarcam mais que uma linha

Em colunas de referência

Em linhas de referência

4.4.2. Identificar células que abarcam mais que uma coluna

4.5. A construção da *template*

4.5.1. Oportunidades para a extracção em contexto contabilístico

4.5.2. Ameaças à extracção em contexto contabilístico

4.5.3. A construção da *template*

4.6. Ordenação das tabelas por grau de interesse

4.6.1. Um método apto em qualquer contexto

4.6.2. Aproveitando as oportunidades do contexto específico

4.6.2.1. Hipótese A

4.6.2.2. Hipótese B

4.7. Análise estrutural

4.8. Interpretação

4.8.1. Captação das relações partitivas

Encontrar a hierarquia

Correcção de erros

Comparação com a *template*

4.8.2. Identificação da unidade de medida

4.8.3. Verificação das relações entre tabelas e entre colunas da mesma tabela

5. Conclusão

O objectivo que pretendemos atingir nesta tese foi encontrar uma forma de compilar determinados itens de informação contabilística, que são transmitidos em tabelas nos relatórios e contas das empresas portuguesas. Estes documentos visam descrever o desenrolar da actividade da instituição no período em causa.

O funcionamento do método que desenhamos para atingir este objectivo, esquematizado na figura abaixo, toma um documento com diversos tipos de elementos gráficos, detecta nele as tabelas, segmenta-as em células, reconhece as que caracterizam o conteúdo da tabela e relaciona-as com os dados transmitidos, interpretando depois o panorama assim formado, para identificar os itens que à partida tinham sido definidos como importantes e extraíndo-os para um depósito de informação agregada.

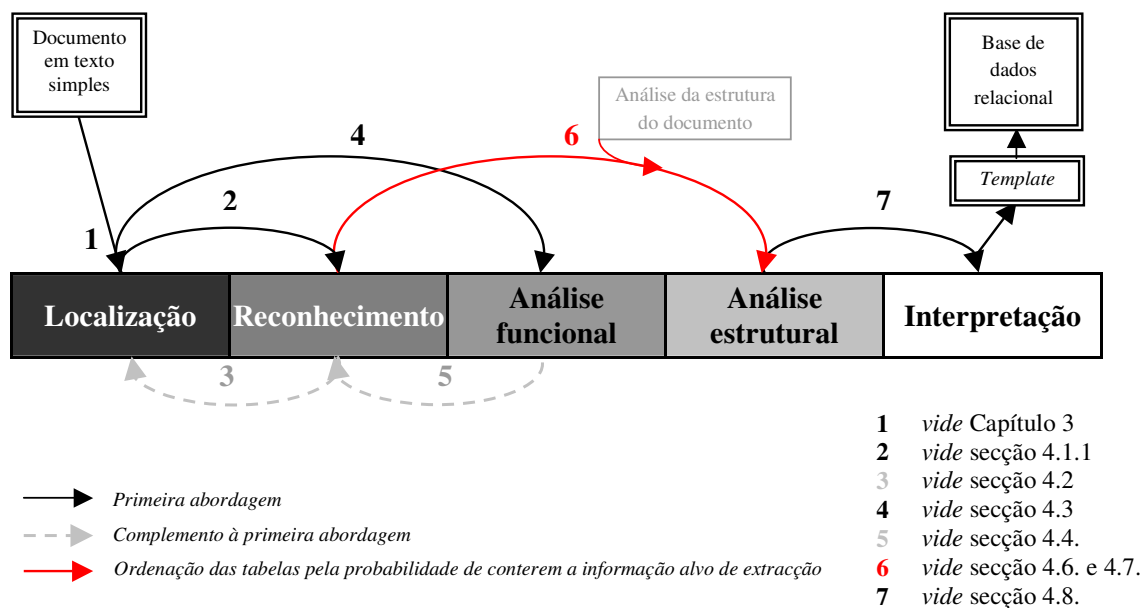


Figura 18. Esquema das relações de *input-output* entre as diferentes fases de resolução do problema.

Como podemos verificar, o sistema não funciona de modo linear: as diferentes tarefas tradicionalmente ligadas ao estudo computacional de tabelas não estão estanques, antes o conhecimento reunido em cada fase é utilizado para aperfeiçoar o desempenho das

tarefas anteriores, aumentando a confiança das decisões tomadas e restringindo a aplicação de certas tarefas mais trabalhosas aos casos em que existe maior probabilidade de a porção de relatório a tratar conter a informação que queremos extrair.

Nesta tese, pudemos percorrer as diferentes fases do trabalho para uma primeira tarefa, a de localizar porções do documento que têm elevada probabilidade de conterem as tabelas cujo conteúdo queremos extrair, classificando cada linha como tendo ou não características de tabela. Porque uma linha isolada não é uma tabela, a classificação não pode depender exclusivamente de factores que digam respeito à linha. Assim, o método de localização utilizado recorre a factores descritivos do contexto em que a linha está inserida: o documento como um todo, a página em causa e as linhas da sua vizinhança. Além disso, o método inclui alguns parâmetros que são manipuláveis pelo utilizador para maximizar a probabilidade de serem identificados determinados tipos de tabelas. Os valores *default* destes parâmetros estão afinados para detectar tabelas grandes ou que partilhem a página com outras tabelas pequenas e conseguem, em 17 dos 19 documentos, descartar mais de 74,4% das linhas mantendo 99,6% das linhas das tabelas com informação passível de extracção. Nos restantes dois relatórios, por manipulação dos parâmetros *default* é possível descartar 62,3% dos registos, conservando 100% dos desejados. Em três relatórios utilizados exclusivamente para teste, os valores *default* do método conseguiram descartar 73,6% das linhas mantendo 99,4% de linhas que contêm informação que queremos extrair.

Na primeira abordagem ao reconhecimento, o método proposto foi implementado e testado numa escala reduzida. Nas fases seguintes, definiram-se as estratégias de resolução dos diversos sub-problemas a que o método tem de dar resposta. Foi sempre nosso objectivo que os métodos utilizados para tratar tabelas fossem adaptáveis a qualquer contexto e não dependessem, tanto quanto possível, do contexto específico dos relatórios e contas. Porém, a fase de interpretação da informação dificilmente é independente do seu contexto, nem se deseja que o seja, quando o contexto pode acrescentar restrições que permitem medir a confiança nos resultados extraídos. O que é desejável, isso sim, é que o método permita acomodar facilmente estas informações contextuais, tal que, não sendo geral na fase de interpretação, seja generalizável. Isto é

exequível pelo modo como se insere no método a informação que é objectivo de extracção, que deve incorporar as relações existentes entre as variáveis que se deseja encontrar.

Contribuições

Como principais contribuições desta tese, temos que o método proposto:

 **Utiliza um *input* maximamente simples:** funciona a partir de ASCII, em que se pode transformar qualquer forma mais complexa, como documentos em Word, XML, Acrobat ou documentos em imagem (com a aplicação de um *software* OCR) ou mesmo digitalizações de documentos em papel. Não necessita de qualquer *markup* para cumprir o(s) seu(s) objectivo(s), partindo de um *input* maximamente pobre e encontrando nele indícios concretos e mensuráveis que o guiam até à extracção da informação pretendida.

 **Adapta-se ao estilo gráfico do documento alvo:** a primeira parte da localização de tabelas em documentos utiliza indícios que tomam em conta características mensuráveis nas quais se vertem os diferentes estilos de organização de elementos gráficos e texto em páginas. A transformação desses estilos em medidas claras, que afectam o funcionamento do método, permite maximizar a qualidade dos resultados obtidos em cada documento: naqueles que têm um estilo mais simples, poder-se-á maximizar a economia (uma vez que altas sensibilidades estão basicamente asseguradas), enquanto nos que têm um estilo mais diverso funcionar-se-á de modo mais cauteloso, procurando maximizar a sensibilidade e estando-se disposto a abdicar de mais economia. Esta capacidade garante ao método flexibilidade e robustez em se adaptar a estilos gráficos em constante mutação.

 **É parametrizável:** o utilizador pode, através da manipulação de alguns parâmetros, afectar o desempenho do sistema, para maximizar a probabilidade de encontrar determinado tipo de tabela.

 **Baseia-se em diversos indícios:** integra nas decisões a tomar elementos de natureza gráfica, mas também linguagem natural e análise da estrutura do documento.

 **Reduz sequencialmente o espaço de procura:** procura, em cada fase do método global, ter em conta o máximo de indícios disponíveis para caracterizar e decidir acerca dos elementos gráficos que estão em análise a cada momento,

restringindo gradualmente o espaço de procura das análises que se seguem. Assim, aquando da análise das linhas (passo 1 da Figura 18), procura-se recolher um conjunto de elementos que as caracterizam e permitem decidir quais as que, apenas com base nas características que a elas se referem, não são de interesse. Nas que passam esta fase, serão encontrados elementos gráficos mais avançados, no caso células (passo 2), e com base nas características que descrevem estas, relativas ao conteúdo e aspectos gráficos, leva-se mais longe o processo de escolher as que têm maior probabilidade de conter informação interessante (passo 3). Destas células procurar-se-á em seguida encontrar elementos mais avançados, nomeadamente distinguir os conjuntos que caracterizam dados dos que contêm dados (passo 4), utilizando-se depois esta nova informação na resegmentação das células e na redelimitação do início e fim das tabelas (passo 5), e assim sucessivamente, até que restem apenas os itens alvo de extracção. O objectivo é conseguir máxima confiança nos resultados de cada fase e finalmente do método, com o mínimo de recursos, não desperdiçando nenhuma fonte de informação que, a dado momento, seja possível captar, de modo a poupar os recursos despendidos na análise das fases seguintes e melhorar as decisões das fases anteriores.



É sensível ao contexto em que opera: incorpora, de modo generalizável a outros contextos, as restrições do contexto específico em que opera. Estas restrições, que permitem medir a confiança dos resultados a que finalmente se chega, incluem a proximidade relativa da informação que se procura, a expectativa dos itens que deverão surgir numa mesma tabela, a ordem em que possivelmente se espera que surjam e as relações aritméticas entre itens de informação numérica, entre outros.

Trabalho futuro

No tempo disponível à elaboração desta tese, foi possível implementar e testar apenas uma parte do método que permitirá alcançar o objectivo proposto. Este foi globalmente planeado, nomeadamente no modo como as diferentes fases podem funcionar e se devem encadear, e de futuro irá certamente sofrer ajustamentos, à medida que a experimentação for elucidando os aspectos a melhorar. Assim, de futuro, há que:

- adaptar o método de localização à captação de tabelas estreitas (com poucos espaços brancos relativamente ao normal do documento) e que utilizem como delimitador de células sequências de símbolos contíguos diferentes de espaço;
- medir o desempenho do algoritmo de delimitação de células (apêndice 2);
- testar em ambiente global a capacidade de calcular o número total de colunas de cada tabela e de atribuição das células às colunas assim identificadas;
- utilizar a informação recolhida sobre cada célula / conjunto de células para refinar o processo de localização de tabelas;
- desenvolver e implementar um método para caracterizar cada linha e coluna das tabelas encontradas como contendo ou descrevendo dados;
- implementar e testar o método de identificação e segmentação das células abarcadoras;
- desenvolver um modo para identificar nomes de organizações;
- identificar títulos no documento, que incorporem a sua estrutura;
- integrar os títulos e algumas características das tabelas encontradas para medir a probabilidade de conterem a informação que se deseja extrair;
- testar em ambiente global a capacidade de identificar uma hierarquia dos conceitos apresentados em dada tabela;
- comparar a informação contida nas tabelas em análise com os itens de informação que desejamos extrair;
- testar a capacidade de reconstruir os caminhos de leitura das tabelas;
- testar as relações aritméticas que devem existir entre os itens a extrair e utilizar a sua verificação (ou não verificação) como indício para ordenar a revisão das decisões que conduziram aos resultados e/ou alertar o utilizador para eventuais erros na informação de base;
- testar em ambiente global a capacidade de identificar correctamente as unidades de medida utilizadas em cada item da tabela;
- preencher a *template* com os elementos extraídos do documento em causa.

Apêndice 1

Testes de hipóteses

1.1. O número de espaços interiores é independente da pertença a uma tabela?

Pretendemos saber se a variável “Número de espaços interiores por linha” tem capacidade de distinguir entre linhas pertencentes a tabelas e linhas que não estejam incluídas neste tipo de elemento gráfico. Se estivéssemos diante de uma variável contínua, bastava-nos para o efeito aplicar um teste às diferenças de médias, mas temos uma variável discreta, que só assume valores pares, não ocupando mais de 100 categorias distintas. Além disso, as amostras que pretenderíamos opor, o conjunto das linhas de tabelas versus o conjunto de linhas sem tabelas, não são independentes, pois estamos diante de diferentes classificações possíveis obtidas para um conjunto de linhas seleccionados.

Assim, para testarmos a hipótese em causa, utilizaremos o Teste do Qui-quadrado. Se não rejeitarmos a hipótese nula, isto quer dizer que o comportamento da variável é distinto entre os dois subconjuntos em oposição. Para podermos aplicar este teste, é necessário garantir que não mais de 20% das probabilidades esperadas é menor que 5 e nenhuma é menor que 1, pelo que temos de combinar a variável Espaços em intervalos mais amplos. Optamos por criar duas classes extremas, 0 e ≥ 124 , e constituir as restantes categorias como 12 intervalos de amplitude constante igual a 10. Assim, sendo $P(A_i)$ a probabilidade de uma linha ser de tipo i , sendo $i = 0, 1$ (não pertence a tabela / pertence a tabela), e $P(B_j)$ a probabilidade de pertencer a um intervalo j , em que $j = 1, \dots, 14$:

Hipótese 0: $P(A_i \cap B_j) = P(A_i) \times P(B_j)$

Hipótese 1: $P(A_i \cap B_j) \neq P(A_i) \times P(B_j)$

As probabilidades são desconhecidas, mas podem ser estimadas do seguinte modo:

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \text{e} \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$$

Sendo e_{ij} a frequência esperada para o conjunto de elementos com tipo i do intervalo j , sob a hipótese nula:

$$\chi^2 = \sum_{i=0}^1 \sum_{j=1}^{14} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx \chi^2(13)$$

Com 0,1% de significância, a região crítica é dada por $]34,53; +\infty[$. A seguir apresenta-se a respectiva tabela de contingência:

Tabela

<i>Espaços</i>	<i>Sim</i>		<i>Não</i>		<i>total</i>
	<i>Frequência absoluta</i>	<i>Frequência esperada</i>	<i>Frequência absoluta</i>	<i>Frequência esperada</i>	
[0]	55.931	47.314	2.975	11.592	58.906
[2;12]	5.931	5.266	625	1.290	6.556
[14;22]	3.599	3.420	659	838	4.258
[24;32]	1.798	2.143	870	525	2.668
[34;42]	1.088	1.713	1.045	420	2.133
[44;52]	831	1.721	1.312	422	2.143
[54;62]	519	1.660	1.548	407	2.067
[64;72]	315	1.671	1.765	409	2.080
[74;82]	184	1.554	1.751	381	1.935
[84;92]	109	1.255	1.454	308	1.563
[94;102]	65	980	1.155	240	1.220
[104;112]	67	734	847	180	914
[114;122]	46	507	585	124	631
[124;+∞[	74	618	695	151	769
	70.557	70.557	17.286	17.286	87.843

Tabela 26. Tabela de contingência entre Tabela e Espaços interiores

A estatística de teste é de 45.249, pertencendo portanto à região crítica, pelo que, com 0,1% de significância e com base nesta amostra, afirma-se que as duas variáveis não são independentes. Assim, pode-se afirmar que a variável tem comportamentos diferentes nos dois grupos.

1.2. Relatórios diferentes apresentam diferentes distribuições de espaços internos?

Para testar se de facto relatórios diferentes têm diferentes distribuições de espaços interiores, decidimos utilizar o teste de Kruskal and Wallis. Com efeito, todos os pressupostos que devem cumprir-se para que o teste seja aplicável, Conover (1980), estão aqui presentes: existe independência mútua entre as diferentes amostras, pois trata-se de relatórios distintos; a unidade de medida é ordinal; as amostras são mais que aleatórias das suas populações, as amostras são as populações – conhecemos a totalidade dos elementos de cada relatório; ou a função distribuição das k populações é idêntica ou algumas populações tendem a ter valores superiores a outras.

Hipótese 0: as populações são idênticas $\Leftrightarrow F_i(x) = F_j(x), \forall x, i, j, i = 1 \dots 19$ e $j = 1 \dots 9$ e $i < j$

Hipótese 1: existem populações diferentes $\Leftrightarrow \exists i, j: F_i(x) < F_j(x), i = 1 \dots 19, j = 1 \dots 19, i < j$

A estatística de teste que se utiliza é a seguinte:

$$T = \frac{1}{S^2} \left(\sum_{i=1}^{19} \frac{\left[\sum_{j=1}^{n_i} R(X_{ij}) \right]^2}{n_i} - \frac{N(N+1)^2}{4} \right) \approx \chi^2(18)$$

e em que

$$S^2 = \frac{1}{N-1} \left(\sum_{\text{tudo}} R(X_{ij})^2 - N \frac{(N+1)^2}{4} \right)$$

sendo a região crítica para 0,1% de significância dada por $] 42,31; +\infty [$. A estatística de teste toma valor 2.960, pertencendo claramente à região crítica. Assim, com 0,1% de significância e com base nesta amostra, podemos afirmar que as distribuições não são todas iguais.

Este teste permite ainda um procedimento que identifica as amostras mais divergentes, que mais contribuem para a rejeição da hipótese nula. Exploremos esta sua faceta. Pode-se dizer que duas populações i e j aparentam ser diferentes se se satisfizer a seguinte

desigualdade (sendo t_{crit} definido a partir da distribuição T de Student com $n - k$ graus de liberdade e 1% de significância):

$$\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right| > t_{crit} \left(S^2 \frac{N-1-T}{N-k} \right)^{\frac{1}{2}} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)^{\frac{1}{2}}$$

	Autoindustrial			Brisa		Cimpor		EDP		Galpenergia		Lusomundo	Novabase	Portucel	RAR	Shell	Siemens	Sonae	Sonae
	98	99	00	99	00	00	97	98	99	00	00	99	00	00	00	99	00	99	00
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1																			
2																			
3																			
4																			
5																			
6																			
7																			
8																			
9																			
10																			
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
<i>Semelhanças</i>	2	2	2	3	2	5	10	5	5	5	5	8	7	0	3	2	1	5	6
<i>Páginas</i>	34	43	56	101	87	234	24	44	168	196	107	33	12	103	32	41	38	81	79
<i>Média Espaços</i>	5	4	4	13	18	14	13	20	14	15	20	18	12	9	21	13	16	22	25
<i>DP espaços</i>	17	15	13	28	29	25	22	30	24	29	39	29	28	24	44	29	25	37	40
<i>Peso de tabelas</i>	10	6	9	16	14	14	37	37	22	26	22	31	22	14	27	20	11	27	28

Tabela 27. Matriz de “dissemelhança” entre os relatórios (a azul, os relatórios que não são considerados semelhantes)

No quadro acima os relatórios estão nomeados de acordo com a Tabela 10, apresentada no capítulo 3; a azul, assinalam-se os pares que o teste considerou diferentes e a branco, os semelhantes; os números em baixo representam o total de pares semelhantes que foi possível formar para dado documento, tendo sido possível definir um total de 72 pares semelhantes, de entre 342 que teria sido possível formar (apenas 22,8%).

Observa-se que os relatórios da mesma empresa são geralmente semelhantes, como se acontece com a Autoindustrial e a Sonae e menos com a EDP. O relatório com maior número de semelhanças, o da EDP em 1997, é um dos mais resumidos e consegue parecer semelhante a relatórios extremamente prolixos. Ao que parece, o número de páginas não tem grande capacidade de afectar a distribuição do número de Espaços interiores como um todo, o que fará algum sentido, se tivermos em conta que muitas vezes os relatórios com mais páginas são muitas vezes preenchidos com gráficos e tabelas não exigidas pelo Plano Oficial de Contabilidade. O mesmo pode ser dito do Peso das Tabelas, que foi calculado como sendo o número de linhas pertencentes a tabelas a dividir pelo número total de linhas do relatório.

Evidentemente, a média e o desvio padrão, indicadores de localização e dispersão respectivamente, são bons indicadores de quais os relatórios com distribuição semelhante de acordo com este teste, pois este está construído para detectar precisamente diferenças de localização e dispersão de distribuições.

1.3. As variáveis Tabela e Relatório são independentes?

Ambas as variáveis são discretas, pelo que para testar a hipótese podemos aplicar o Teste do Qui-quadrado às frequências esperadas, se não forem demasiado pequenas.

Hipótese 0: as duas variáveis são independentes $\Leftrightarrow p_{ij} = P(A_i \cap B_j) = P(A_i) \times P(B_j)$

Hipótese 1: as duas variáveis não são independentes $\Leftrightarrow$

$$p_{ij} = P(A_i \cap B_j) \neq P(A_i) \times P(B_j)$$

As probabilidades são desconhecidas, mas podem ser estimadas do seguinte modo:

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \text{e} \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$$

Sendo e_{ij} a frequência esperada para o conjunto de elementos com tabela = i , sendo $i = 1$ ou 0 conforme dada linha pertença a uma tabela ou não, provindos do relatório j , com j a variar entre 1 e 19, sob a hipótese nula:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^{19} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx \chi^2(18)$$

Com 0,1% de significância, a região crítica é dada por $]42,31; +\infty[$. A seguir apresenta-se a respectiva tabela de contingência:

Tabela

ID	<i>Sim</i>		<i>Não</i>		<i>total</i>
	<i>Frequência absoluta</i>	<i>Frequência esperada</i>	<i>Frequência absoluta</i>	<i>Frequência esperada</i>	
1	1.852	1.661	216	407	2.068
2	2.159	1.838	130	451	2.289
3	2.549	2.245	247	551	2.796
4	4.396	4.208	844	1.032	5.240
5	4.712	4.408	777	1.081	5.489
6	13.364	12.435	2.121	3.050	15.485
7	532	680	315	167	847
8	1.582	2.009	920	493	2.502
9	6.535	6.711	1.822	1.646	8.357
10	7.211	7.875	2.596	1.932	9.807
11	5.611	5.749	1.548	1.410	7.159
12	1.505	1.757	683	431	2.188
13	325	334	91	82	416
14	5.305	4.948	856	1.213	6.161
15	1.519	1.674	565	410	2.084
16	1.809	1.811	446	444	2.255
17	1.980	1.793	253	440	2.233
18	3.929	4.316	1.446	1.059	5.375
19	3.667	4.089	1.425	1.003	5.092
total	70.542	70.542	17.301	17.301	87.843

Tabela 28. Tabela de contingência entre Tabela e os distintos relatórios e contas

Nenhuma das frequências absolutas apresenta valor inferior a 5. A estatística de teste é de 2.941, pertencendo portanto à região crítica, pelo que, com 0,1% de significância e com base nesta amostra, afirma-se que as duas variáveis não são independentes.

Apêndice 2

Algoritmo de segmentação de linhas em células

Nomenclatura:

$L \leftarrow$ conjunto de linhas, l , do documento de origem

$O \leftarrow$ conjunto de células que serão o output do programa

$\{N_1 \cup N_2\} \leftarrow$ conjunto de caracteres considerados admissíveis num número

$N_1 \leftarrow \{ '1', '2', '3', '4', '5', '6', '7', '8', '9', '0' \}$

$N_2 \leftarrow \{ '-', '+', '\%', ',', ';', ':', '"', '(', ')' \}$

$c_1, c_2, c_3 \leftarrow$ três caracteres contíguos

$strnv \leftarrow$ conteúdo de uma célula

$strnv \leftarrow$ primeiro caracter do conteúdo da célula seguinte

$n \leftarrow$ primeira estimativa da coluna em que a célula se situa

$sinalizador \leftarrow$ assume valor 0 se a decisão não é de risco e diferente de 0 caso contrário

Algoritmo:

$l \leftarrow$ primeira linha de L

$O \leftarrow \{ \}$

Para cada linha l de L ,

$n \leftarrow 1$

c_1 e $c_2 \leftarrow$ caracter espaço, ' '

$c_3 \leftarrow$ primeiro caracter da linha l

% até se chegar ao fim da linha

Enquanto $c_3 \neq ''$

% verifica se os caracteres podem formar um número

Se $c_3 \in \{N_1 \cup N_2\}$ **e** $numant = 1$ **então**

$numnovo = 1$

senão $numnovo = 0$ **e** $numant = 0$

Fim de se.

% decide como agrupar os três caracteres em análise

Se $c_2 \leq \Rightarrow$ ‘ ’ **e** $c_3 \leq \Rightarrow$ ‘ ’ **e** $c_2 \leq \Rightarrow$ ‘ ’ **e** $c_3 \leq \Rightarrow$ ‘ ’ **então**

% dois caracteres não espaço contíguos => está-se a meio de uma expressão una
strnv = concatenar (strnv, c_3)

Fim de se.

Se $c_1 \leq \Rightarrow$ ‘ ’ **e** $c_1 \leq \Rightarrow$ ‘ ’ **e** [$(c_2 = \text{' ' e } c_3 = \text{' '})$ **ou** $c_3 = \text{' '}$] **então**

% dois espaços após um número ou três num texto => guardar strnv em célula

Se numnovo = 1 **ou** [numnovo = 0 **e** ($c_3 = \text{' '}$ **ou** $c_{3+1} = \text{' '}$)] **então**

% alguns passos de tratamento do strnv: retirar espaços excessivos,
garantir coerência de parêntesis, detectar símbolos de coluna vazia.

Se strnv contiver { ‘+’, ‘-’, ‘(’, ‘)’, ‘%’ }

então substituir ‘ ’ por ‘ ’

Fim de se.

Se numnovo = 1 **e** strnv $\leq \Rightarrow$ ‘(’ **e** strnv contiver ‘(’ **e** strnv não
contiver ‘)’

então strnv = concatenar (strnv, ‘)’)

Fim de se.

Se numnovo = 1 **e** strnv $\leq \Rightarrow$ ‘)’ **e** strnv contiver ‘)’ **e** strnv não
contiver ‘(’

então strnv = concatenar (‘(’, strnv)

Fim de se.

Se strnv = ‘(’

então strprox = ‘(’.

Fim de se.

Se strnv in { ‘|’, ‘-’, ‘*’ }

então $n = n + 1$

Fim de se.

% guarda strnv como uma célula e também as suas características
- valor de numnovo, n, comprimento, posição de início,
sinalizador...)

Se strnv $\notin \{N_1 \cup N_2 \cup \text{'*'} \cup \text{'|'}\}$ **então**

Se $numnovo = 1$ **e** $Formata (strnv, data) = False$ **então**

Se $Formata (strnv, numérico)$

Fim de se.

Fim de se.

Fim de se.

$O \leftarrow O \cup \{strnv\}$

$strnv \leftarrow strprox$

$strprox \leftarrow ''$

$sinalizador \leftarrow 0$

$numant \leftarrow 1$

$n \leftarrow n + 1$

Fim de se.

Fim de se.

Se $[(c_1 = '' \text{ e } c_2 = ')] \text{ ou } c_2 = ' \text{ e } c_3 \leq '' \text{ e } c_3 \leq ' \text{ então}$

% dois espaços antes de um caracter não espaço => talvez o início de strnv

Se $strnv \leftarrow ''$ **então**

$strnv \leftarrow \text{concatenar}(strnv, c_3)$

senão $strnv \leftarrow \text{concatenar}(strnv, ', c_3)$

Fim de se.

Fim de se.

Se $c_1 \leq '' \text{ e } c_1 \leq ' \text{ e } c_2 = '' \text{ e } c_3 \leq '' \text{ e } c_3 \leq ' \text{ então}$

*% um espaço entre dois caracteres de um número => ou é separador de milhar
ou divide duas células (excepto quando c_1 seja ponto ou vírgula), caso em que é
lapso e há que excluí-lo; em texto, há que tratá-lo como um caracter normal*

Se $numnovo = 0$ **então**

$strnv = \text{concatenar}(strnv, ', c_3)$

senão se $numnovo = 1 \text{ e } (c_1 = '.' \text{ ou } c_1 = ',) \text{ então}$

$strnv = \text{concatenar}(strnv, c_3)$

senão se $c_{3+3} \in \{N_2 \cup '*'\} \setminus \{', ', ', '\} \text{ ou } (c_{3+3} = '' \text{ e } c_{3+2} \leq '') \text{ ou}$

$(c_{3+3} \in \{', ', '\} \text{ e } [c_{3+5} \in \{', ', '\} \text{ ou } c_{3+6} \in \{', ', '\}]) \text{ então}$

% considera-se que o espaço funciona como separador de milhar
 $strnv \leftarrow \text{concatenar}(strnv, c_3)$
 $sinalizador \leftarrow 1$
senão
% guarda strnv como uma célula e também as suas características -
valor de numnovo, n, comprimento, posição de início, sinalizador...)
 $O \leftarrow O \cup \{strnv\}$
 $strnv \leftarrow c_3$
 $sinalizador \leftarrow 2$
 $numant \leftarrow 1$
 $n \leftarrow n + 1$
Fim de se.
Fim de se.
 $c_1 \leftarrow c_2$
 $c_1 \leftarrow c_3$
 $c_3 \leftarrow \text{próximo caracter da linha } l$
Fim de enquanto.
 $l \leftarrow \text{próxima linha de L}$
Fim de para.
Fim de algoritmo.

Referências bibliográficas

Abu-Tarif, Asad A. (1998), “Table processing and understanding”, Master of science thesis, Rensselaer Polytechnic Institute, Troy, New York, EUA.

Bento, José e José Fernandes Machado (1999), *Plano Oficial de Contabilidade Explicado*, Portugal: Porto Editora.

Cameron, James P. (1989), “A cognitive model for tabular editing”, Technical report OSU-CISRC6/89-TR26, Computer and Information Science Research Center, Ohio State University, Ohio, EUA.

Chang, George, Marcus J. Healey, James A.M. McHugh e Jason T.L. Wang (2000), *Mining the World Wide Web – an information search approach*, Boston / Dordrecht / Reino Unido: Kluwer Academic Publishers.

Chen, Hsin-His, Shih-Chung Tsai e Jin-Hi Tsai (2000), “Mining Tables from Large Scale HTML Texts”, in *18th International Conference on Computational Linguistics (COLING)*, pp.166-172, Saarbrücken, Alemanha.

Conover, W. J. (1980), *Practical nonparametric statistics*, EUA: John Wiley & Sons, Inc.

Constantino, Marco (1997), “Financial Information Extraction using pre-defined and user-definable Templates in the LOLITA System”, PhD Thesis, University of Durham, Reino Unido.

Document Analysis Systems V (2002), 5th International Workshop, DAS 2002, Princeton, NY, EUA (<http://www.research.avayalabs.com/DAS02/>).

Document recognition and retrieval VIII (2001), San Jose, California, EUA (<http://www.spie.org/web/meetings/programs/pw01/confs/4307.html>).

Douglas, Shona, Matthew Hurst e David Quinn (1995), “Using Natural Language Processing for *Identifying* and Interpreting Tables in Plain Text”, in *Fourth Symposium on Document Analysis and Information Retrieval*, pp. 535-545, Nevada, EUA.

EAGLES Lexicon Interest Group (1998), “Preliminary Recommendations on Semantic Encoding”, Interim report, EAGLES - Expert Advisory Group on Language Engineering Standards, Pisa, Itália (disponível em <http://www.ilc.pi.cnr.it/EAGLES96/rep2/rep2.html>).

Ferguson, Don (1997), “Parsing Financial Statements Efficiently and Accurately Using C and Prolog”, in *Practical Applications of Prolog Conference '97*, Londres, Reino Unido.

Green, Edward e M. Krishnamoorthy (1995a), “Table Image Understanding”, in *Information Systems Analysis and Synthesis, Proceedings of the International Conference on Information Systems, Analysis and Synthesis, ISAS'96*, EUA.

Green, Edward e M. Krishnamoorthy (1995b), “Model-based of Printed Tables”, in *Proceeding of International Conference of Document Analysis and Recognition 95, ICDAR*, pp. 214-217, Montréal, Canada.

Handley, John C. (2001), “Table analysis for multiline cell identification”, in *Proceedings of Document Recognition and Retrieval VIII*, SPIE, San Jose, EUA.

Hobbs, J. (1993), “The generic information extraction system”, in *Fifth Messages Understanding conference* (MUC-5), Morgan Kaufmann Publishers, pp. 87-91, Baltimore, Maryland, EUA.

Hu, J., R Kashi, D. Lopresti e G. Wilfong (2000), “Medium-Independent Table Detection”, in *Proceedings of Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging)*, volume 3967, pp. 291-302, EUA.

Hu, J., R Kashi, D. Lopresti e G. Wilfong (1999), “Table detection across multiple media”, in *International Workshop on Document Layout Interpretation and Its Applications*, Bangalor, India.

Hurst, Mathew e Shona Douglas (1997), “Layout and Language: Preliminary investigations in recognizing the structure of tables”, in *Proceedings of ICDAR’97*, PP. 1043-1047, Ulm, Alemanha.

Hurst, Mathew e Shona Douglas (1997), “Layout and Language: Preliminary experiments in assigning logical structure”, in *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 217-220, Reino Unido.

Hurst (2000), “The interpretation of tables in texts”, PhD. Thesis, School of Cognitive Science, Informatics, The University of Edinburgh, Reino Unido.

Hurst, Mathew e Tetsuya Nasukawa (2000), “Layout and Language: Integrating Spatial and Linguistic Knowledge for Layout Understanding Tasks”, in *Proceedings of the 18th International Conference on Computational Linguistics, ICCL*, Saarbruecken, Alemanha.

Hurst, Mathew (2001), “Layout and Language: An Efficient Algorithm for Text Block Detection based on Spatial and Linguistic Evidence”, in *Document Recognition and Retrieval VIII, SPIE Proceedings*, Volume 4307, pp. 55-67, San Jose, CA, EUA.

Kieninger, Thomas e Andreas Dengel (1998), “A paper-to-html table converting system”, in *Proceedings of Document Analysis Systems, (DAS) ’98*, Nagano, Japan.

Kieninger, Thomas (1998), “Table Structure Recognition Based on Robust Block Segmentation”, in *Proceedings of IS&T/SPIE's 10th Annual Symposium Electronic Imaging '98: Science & Technology - Document Recognition V*, San Jose, California, EUA.

Klink, Stefan, Andreas Dengel, Thomas Kienenger (2000), “Document structure analysis based on layout and textual features”, in *Proceedings of Fourth IAPR International Workshop on Document Analysis Systems*, DAS2000, pp. 99-111, Rio de Janeiro, Brasil.

Kornfeld, William e John Wattecamps (1998), “Automatically locating, extracting and analyzing tabular data”, in *Proceedings of the 21st annual international ACM SIGIR conference*, SIGIR '98, pp. 347-349, Melbourne, Australia.

Mendes de Oliveira, M, Álvaro Aguiar, Armindo Carvalho, Francisco V. Martins, Victor Mendes e Pedro Portugal (1997), *Econometria: exercícios*, Portugal: McGraw Hill.

Ng, Hwee Tou, Chung Yong Lim e Jessica Li Teng Koo (1999), “Learning to recognize tables in free text”, in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp.443-450, Maryland, EUA.

Pyreddy, Pallavi e W. Bruce Croft (1997), “A System for Retrieval in Text Tables”, Technical report 105, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, Massachusetts, EUA.

Ribeiro, Antonio, Gabriel Pereira Lopes e João Mexia (2000), “Extracting Equivalents from Aligned Parallel Texts: Comparison of Measures Similarity”, in *IBERAMIA-SBIA 2000*, LNAI 1952, M.C. Monard and J.S. Sichman (Eds.), pp. 339-349, Berlin, Alemanha.

Riloff, E e W. Lehnert (1994), “Information Extraction as a basis for high-precision text classification”, in *ACM Transactions on Information Systems*, Vol. 12, N°3, pp. 296-333.

Rus, Daniela e Kristen Summers (1994), “Using White Space for Automated Document Structuring”, in *Workshop on the Principles of Document Processing*, Seeheim, Alemanha.

Science Applications International Corporation (1997), “An Automated Conversion of Structured Documents into SGML” Technical Report, Distributed Object Computation Testbed (DOCT), San Diego Supercomputer Center (disponível em <http://www.sdsc.edu/DOCT/Publications/a3-3/a3-3.html>)

Shamillian Jonh H, Henry S. Baird e Thomas L. Wood (1997), “A retargetable table reader”, in *Proceedings of the IAPR 97 International Conference on Document Analysis and Recognition 97*, pp. 448-453, Ulm, Alemanha.

Thompson, Marcy (1996), “A tables manifesto”, in *Proceedings of SGML Europe 1996*, pp. 151-153, Munique, Alemanha.

Tersteegen, Wolfgang e Claudia Wenzel (1998), “Scantab: Table recognition by reference tables”, in *Third International Association for Pattern Recognition Workshop on Document Analysis Systems*, DAS 98, Nagano, Japão.

Tupaj, Scott, Zhongwen Shi, C. Hwa Chang e Hassan Alam (1996), “Extracting Tabular Information From Text Files”, EECS Department, Tufts University, Medford, EUA (disponível em <http://www.ee.tufts.edu/hchang/paper1.ps>).

XBRL.org (2000), XBRL Financial Statement Viewer (disponível em <http://www.xbrl.org/us/gaap/ci/2000-07-31/Sample/Viewer.htm>).

Wakao, Takahiro, Robert Gaizauskas, Yorick Wilks (1996), “Evaluation of an algorithm for the recognition and classification of proper names”, in *Proceedings of the 16th International Conference on Computational Linguistics*, COLING96, Copenhaga, pp. 418-423.

Web Document Analysis (2001), First International Workshop (WDA2001), Seattle, Washington, EUA (<http://www.csc.liv.ac.uk/~wda2001/>).

Wikipedia, the free encyclopaedia (disponível em <http://wikipedia.org/>).

Anexo 1

Distribuição global de espaços

Para averiguarmos o comportamento da variável “Número de espaços interiores por linha”, decidimos traçar o respectivo histograma, pintando de cores distintas os elementos que queremos distinguir. Nos gráficos seguintes, distinguem-se o comportamento das linhas que não pertencem a tabelas (ref 0), das que pertencem a tabelas pouco interessantes para nós, que nos é indiferente localizar ou não (ref 1), e das que correspondem a tabelas que contêm informação alvo de extracção (ref 2).

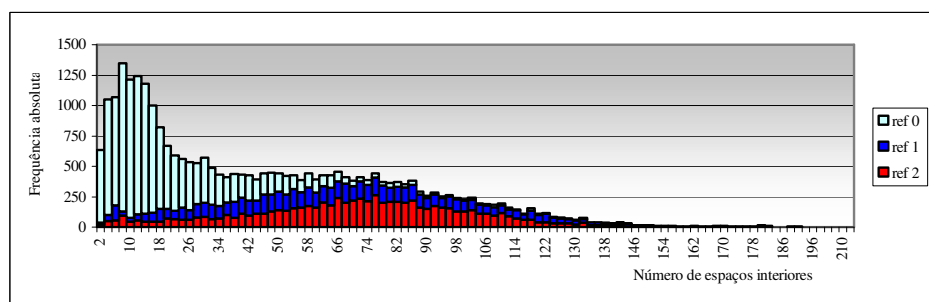


Figura 19. Distribuição de espaços interiores no total da amostra

O gráfico acima não inclui o conjunto de linhas com zero espaços interiores, que, sendo 58.906 registos, tornaria o demais da distribuição ilegível. Neste subconjunto, a proporção de cada uma das classes pode ser visualizada no gráfico adjacente.

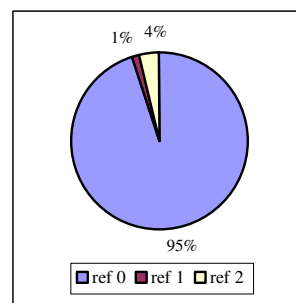


Figura 20. Proporção de tipos quando não há espaços interiores

A evidência empírica parece comprovar a noção que a variável é de facto capaz de diferenciar as tabelas (aliás, se se tivesse de adivinhar, dir-se-ia que em ref 0 temos uma distribuição de tipo F de Snedecor e no seu complementar uma distribuição normal). A variável não parece ser capaz de distinguir entre tabelas interessantes (ref 2) e normais (ref 1).

Anexo 2

Distribuição de espaços nos documentos de origem

Apresentamos a seguir os histogramas da distribuição dos espaços interiores nos diferentes relatórios e contas que compõem o conjunto de treino (dos quais, pelo mesmo motivo que no anexo anterior, excluem as linhas com zero na variável).

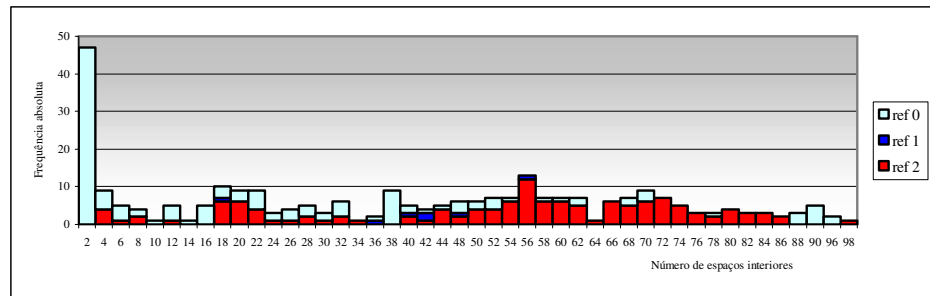


Figura 21. Distribuição de espaços brancos interiores na Autoindustrial 1998

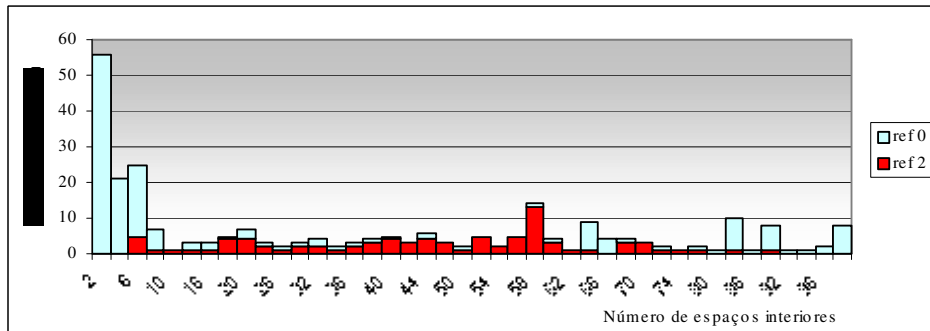


Figura 22. Distribuição de espaços brancos interiores na Autoindustrial 1999

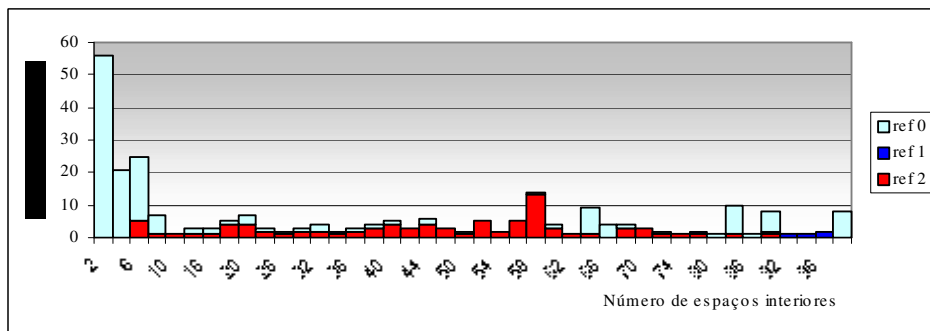


Figura 23. Distribuição de espaços brancos interiores na Autoindustrial 2000

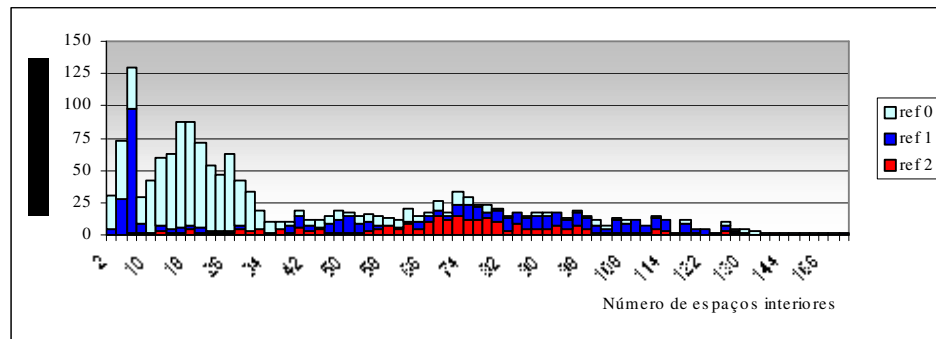


Figura 24. Distribuição de espaços brancos interiores na Brisa 1999

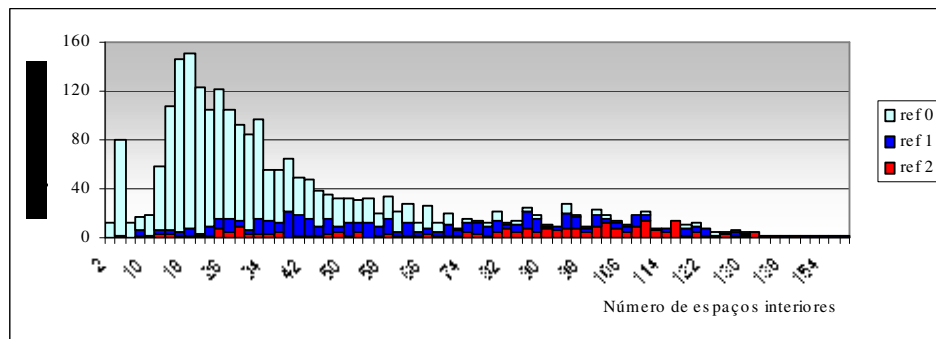


Figura 25. Distribuição de espaços brancos interiores na Brisa 2000

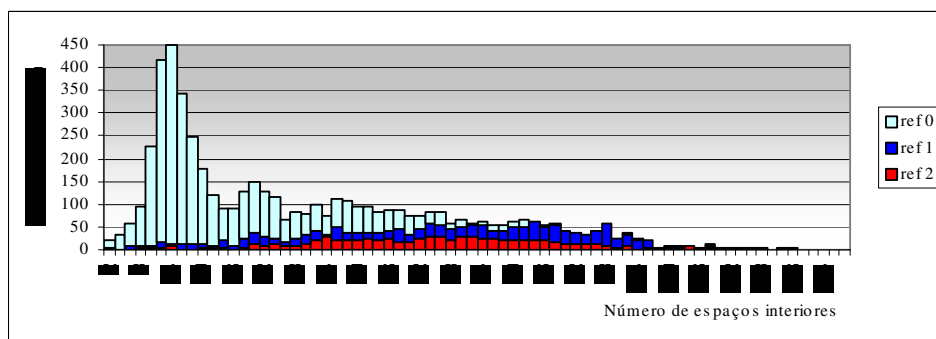


Figura 26. Distribuição de espaços brancos interiores na Cimpor 2000

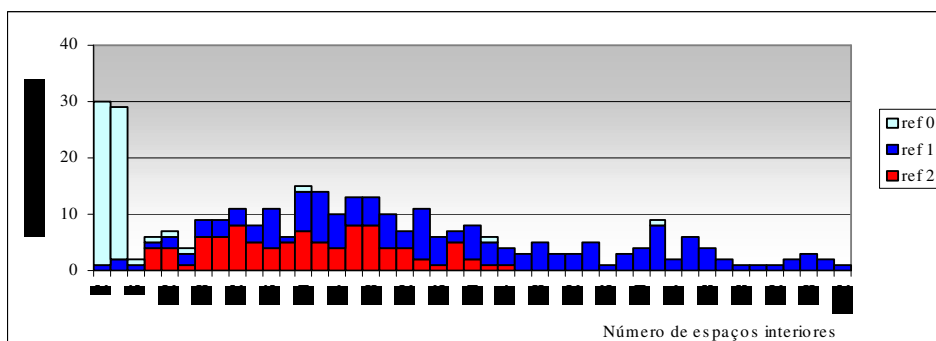


Figura 27. Distribuição de espaços brancos interiores na EDP 1997

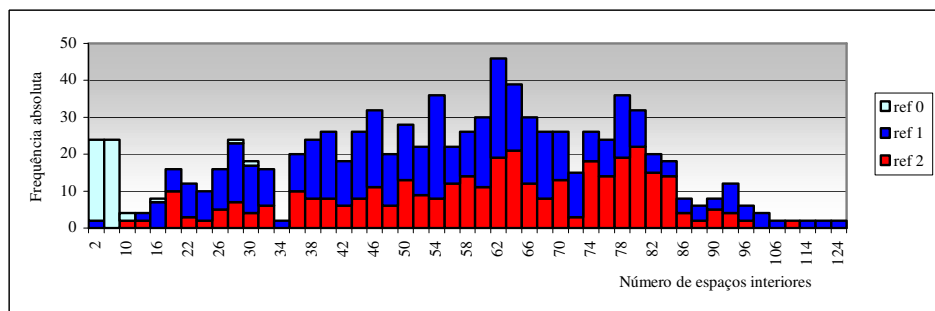


Figura 28. Distribuição de espaços brancos interiores na EDP 1998

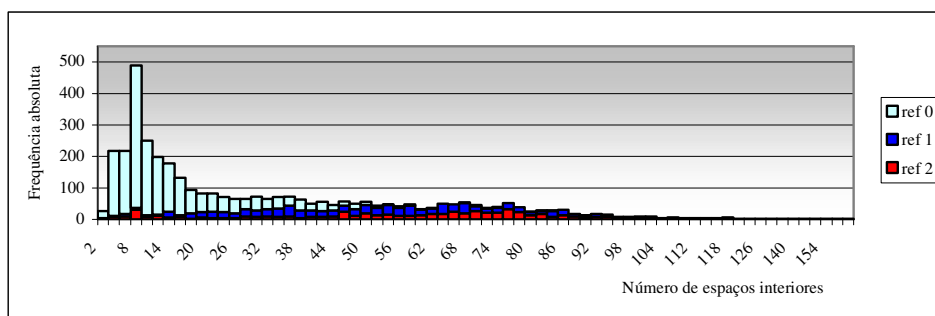


Figura 29. Distribuição de espaços brancos interiores na EDP 1999

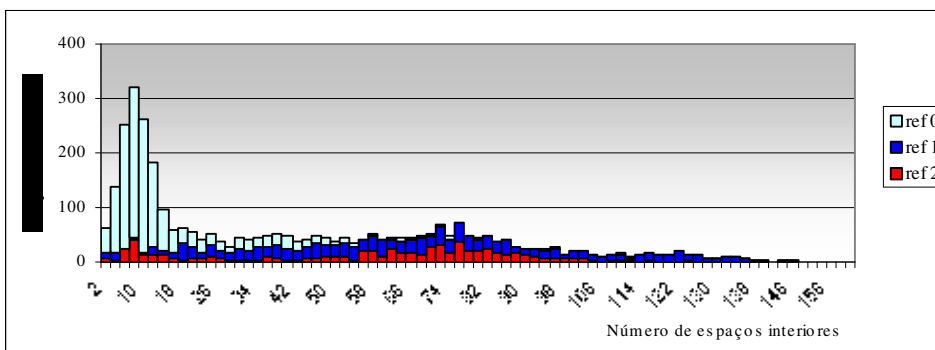


Figura 30. Distribuição de espaços brancos interiores na EDP 2000

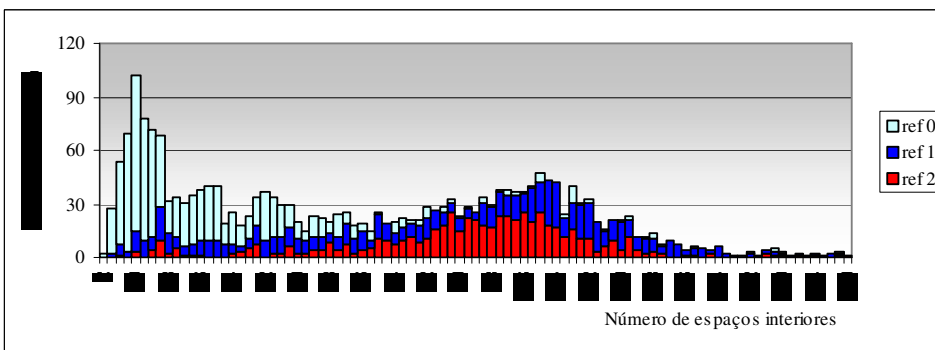


Figura 31. Distribuição de espaços brancos interiores na Galpenergia 2000

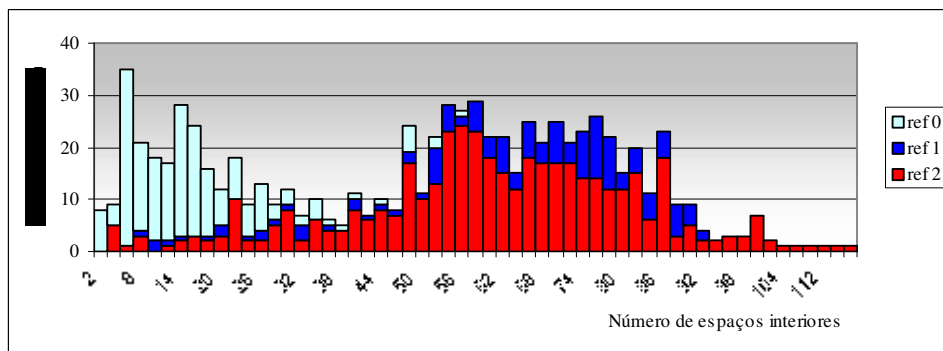


Figura 32. Distribuição de espaços brancos interiores na Lusomundo 1999

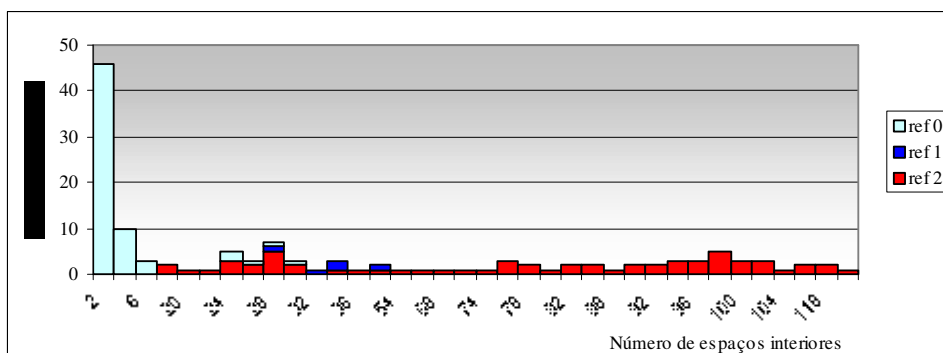


Figura 33. Distribuição de espaços brancos interiores na Novabase 2000

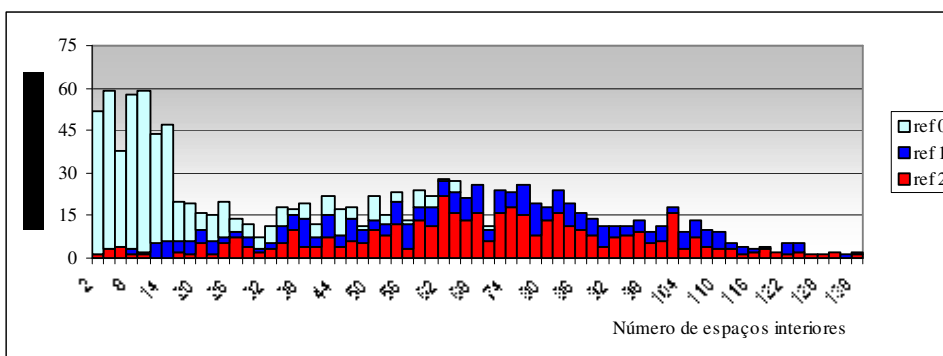


Figura 34. Distribuição de espaços brancos interiores na Portucel 2000

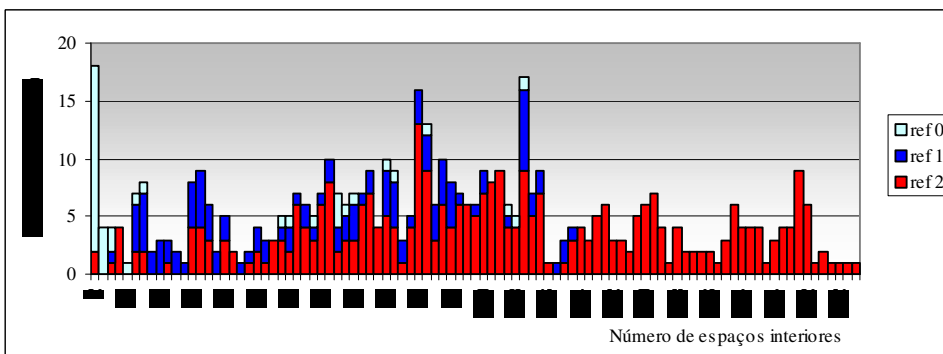


Figura 35. Distribuição de espaços brancos interiores na RAR 2000

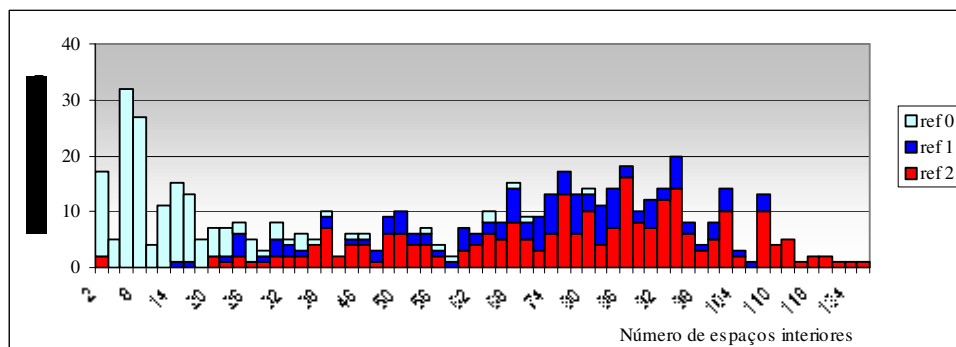


Figura 36. Distribuição de espaços brancos interiores na Shell 1999

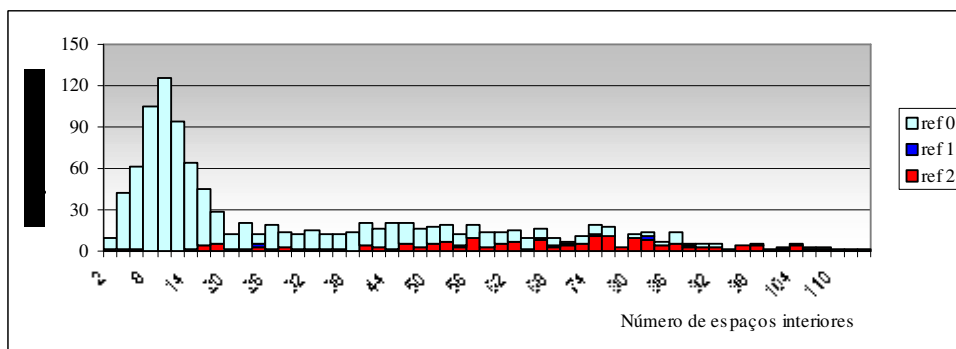


Figura 37. Distribuição de espaços brancos interiores na Siemens 1999

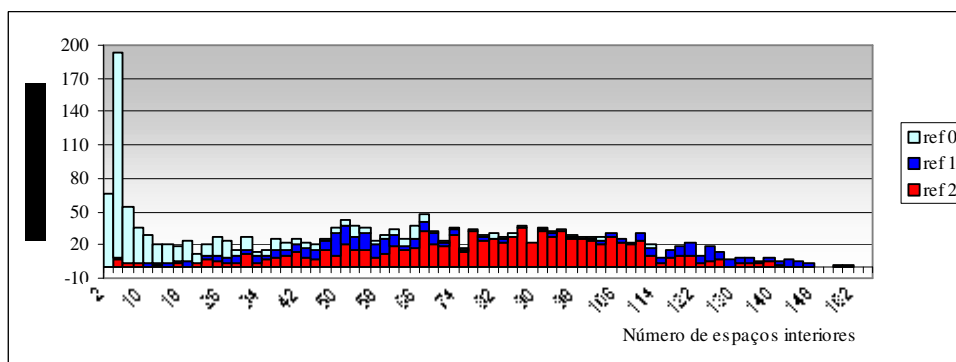


Figura 38. Distribuição de espaços brancos interiores na Sonae 1999

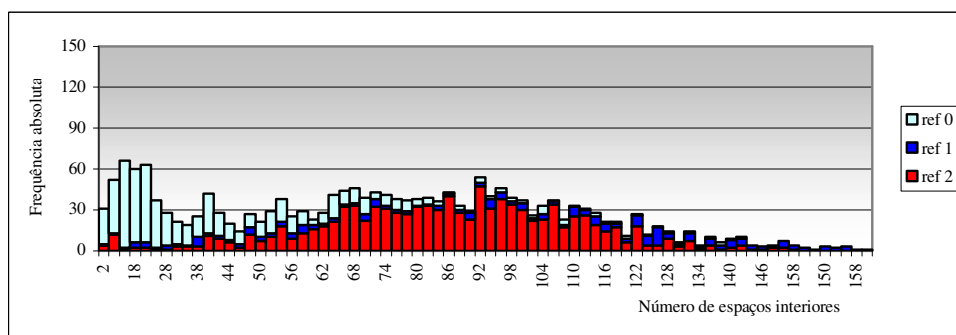


Figura 39. Distribuição de espaços brancos interiores na Sonae 2000

Há algumas constatações empíricas interessantes a fazer:

-  para qualquer dos relatórios, mantém-se a constatação de que os dois grupos – o das tabelas e o de não-tabelas - têm localizações diferentes.
-  a variável apresenta comportamento similar nas tabelas interessantes (ref 2) e nas normais (ref 1).
-  o relatório da Cimpor 2000 é o maior dos que constam da amostra, provindo dele 17% dos registos, sendo também, naturalmente, o que tem um histograma mais parecido com o do Anexo anterior.
-  relatórios da mesma empresa em dois anos distintos têm histogramas relativamente parecidos (excepção feita à EDP, que foi privatizada em 1999, altura a partir da qual a riqueza de conteúdo do seu relatório passou a ser muito mais importante para ela).
-  entre empresas diferentes, o aspecto das distribuições dos dois grupos pode ser muito distinto, em termos, por exemplo, do peso relativo dos dois grupos, de localização, de dispersão e de assimetria, conforme se verifica na tabela seguinte, em que os relatórios estão ordenados pela respectivas médias.

RC_ID	Número de linhas	Média	Desvio padrão	Coefficiente de variação	Peso de tabelas	Skewness	Máximo fora 0
3	358	29	26	0,88	9%	0,50	2
9	3.688	31	27	0,87	22%	1,11	8
17	1.108	32	27	0,86	11%	0,95	10
2	256	34	32	0,96	6%	0,63	2
7	317	35	24	0,70	37%	0,56	2
13	131	37	40	1,08	22%	0,70	2
1	276	39	28	0,72	10%	0,13	2
6	5.325	40	29	0,72	14%	0,83	14
5	2.398	41	31	0,75	14%	1,23	18
10	3.430	42	36	0,86	26%	0,78	8
4	1.638	43	36	0,84	16%	0,87	6
14	1.200	47	35	0,73	14%	0,34	4
12	790	50	28	0,56	31%	-0,15	6
8	900	54	23	0,43	37%	-0,29	62
16	536	57	35	0,62	20%	-0,13	6
18	1.961	59	40	0,68	27%	0,17	10
11	2.225	66	45	0,68	22%	0,26	10
19	1.918	66	39	0,59	28%	-0,07	16
15	482	89	50	0,56	27%	0,16	2

Tabela 29. Algumas medidas de estatística descritiva da distribuição do Número de espaços interiores por linha, nos diversos relatórios e contas

Anexo 3

Avaliação em contexto específico

[Anexos.xls](#)

Anexo 4

Avaliação em contexto genérico

[Anexos.xls](#)