PersonalNews: A Nossa Vida de Relance…

Bruno Antunes
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
bruno.r.t.antunes@ist.utl.pt

Tiago Guerreiro
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
tjvg@immi.inesc.pt

Daniel Gonçalves
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
daniel.goncalves@inesc-id.pt

Sumário

Com o aumento da quantidade de informação existente nos computadores pessoais, existe cada vez mais a necessidade de desenvolver novas formas de gerir essa informação. Apresentamos uma interface em que é possível visualizar e navegar de forma eficaz na informação pessoal contida num determinado computador pessoal, possibilitando a descoberta de padrões relevantes nessa informação. Isto é feito criando um jornal pessoal *do utilizador, em que várias notícias descrevem os desenvolvimentos nos temas que mais interessaram o utilizador num dado intervalo de tempo. Essas notícias, automaticamente inferidas a partir dos documentos do utilizador, são dispostas no jornal de acordo com a sua importância relativa, dando ao utilizador de uma ideia dos temas que mais o interessaram e permite facilmente encontrar os documentos com eles relacionados.*

Palavras-chave

Gestão de Informação pessoal, visualização, sumários, notícias.

1. INTRODUÇÃO

Actualmente, os computadores pessoais contém cada vez mais informação. Numa determinada altura, um utilizador pode estar a trabalhar, criar e alterar vários documentos de diversas áreas e sobre diferentes temas. Existe assim uma grande quantidade de informação pessoal, que reflecte os nossos interesses e actividades, dispersa por vários locais. Por exemplo, os emails e documentos referentes a uma mesma tarefa são geridos de formas díspares e nem sempre interoperáveis. Em particular, não é possível fazê-lo de modo semanticamente relevante para o utilizador, agrupando a informação de acordo com o contexto, actividades e interesses.

Para o conseguir, há dois obstáculos a transpor: indexar a informação pessoal e apresentar essa informação ao utilizador para que a sua utilização seja relevante e eficaz. Se em relação ao primeiro problema, surgem cada vez mais soluções, a resolução do segundo não é igualmente simples. Mecanísmos de indexação e pesquisa de *desktop*, como é o caso do Google Desktop, fornecem apenas um modo de interacção básico (pesquisa por palavras-chave), permitindo ao utilizador recuperar documentos individuais mas não dando uma visão de conjunto.

Propomos uma abordagem que possibilita aos utilizadores visualizarem a sua informação pessoal de modo a revelar de forma imediata o contexto e interesses do utilizador num dado intervalo temporal. Tal é conseguido pela criação automática de grupos de documentos semanticamente relacionados de acordo com os vários temas de interesse para o utilizador. Uma vez identificados os temas, estes são apresentados ao utilizador sob a forma de notícias num *jornal pessoal*. Esta é uma metáfora amplamente conhecida e fácil de compreender, adaptando-se à grande quantidade de informação a mostrar. Cada "edição" do jornal reporta-se a um determinado período temporal, reflectindo as notícias os interesses do utilizador nesse período. O texto das notícias é gerado automaticamente a partir dos documentos em cada tema, podendo estes ser recuperados directamente a partir do PersonalNews.

O trabalho foi desenvolvido em quatro grandes partes: criação dos temas, extracção dos sumários sobre os vários temas, criação de notícias e a apresentação das notícias ao utilizador. A criação de temas possibilita a identificação das áreas em que o utilizador trabalhou ao longo do tempo. Uma vez indentificados, a extracção de sumários permite obter uma breve descrição de cada um, com base nos documentos neles contidos. Com base nessa descrição é possível criar as noticias que irão figurar no jornal, sendo a geração deste, lidando com o *layout* das várias notícias, e interface de navegação, a última parte.

Na próxima secção iremos abordar o estado da arte e os trabalhos relacionados com o nosso. Na Secção 3 descreveremos a interface do PersonalNews. Em seguida mostraremos como são inferidos os temas a partir dos documentos, e na Secção 5 descreveremos a criação de sumários para os temas e a geração das notícias propriamente dita. Concluiremos, então, apresentando algumas ideias de trabalho futuro na área.

2. TRABALHO RELACIONADO

Nos últimos anos têm surgido variadíssimas aplicações e estudos na área no qual este artigo se insere, a visualização de informação pessoal. Um dos primeiros estudos nesta área foi o *Forgot-me-not* [Lamming94]. Este artigo descreve um dispositivo semelhante a um

PDA em que é representado um histórico da informação pessoal passada, tentando assim ajudar a memória do seu utilizador. Por exemplo, permite lembrar onde se encontra um determinado documento ou o nome de uma determinada pessoa. É no entanto um trabalho relativamente antigo que está aquém do que hoje se pode conseguir, especialmente em termos de interface.

Mais recentemente, o *Themail* [Viégas06] é um sistema em cujo objectivo é criar uma visualização da informação preservada em arquivos e-mail dos seus utilizadores. Permite dar resposta a questões como "de que coisas falo com cada um dos meus contactos e-mail" ou "como é que as minhas conversas diferem das outras pessoas". Assim sendo, o *Themail* mostra a relação entre o utilizador e os seus contactos de e-mail, ao longo do tempo, ao escolher e mostrar as palavras-chave mais relevantes das mensagens trocadas entre estes. A interface é atraente e eficaz. No entanto, está limitado à informação contida nos arquivos do e-mail. Outro problema desta aplicação é tratar todas as mensagens da mesma forma, não tendo em conta a importância relativa destas.

Um outro dispositivo interessante é o *FacetMap* [Smith06], um sistema de visualização interactiva, guiado por *queries*, generalizável para um grande conjunto de informação através da sua meta-informação. O princípio base por detrás do *FacetMap* é mostrar qualquer conjunto de dados da maneira mais eficaz, dado a restrição do tamanho do ecrã, o número de itens e os seus atributos. Embora os autores conseguissem criar um novo paradigma para a pesquisa e navegação na informação, os testes realizados com utilizadores mostraram que ficavam confundidos com a interface e com as várias *facets*, os tópicos usados para organizar a informação (tempo, tipo, autor, etc.). Estes são representados por ovais que podem ser pressionadas para filtrar os resultados das pesquisas.

Por fim, o *Milestones In Time* [Ringel03] descreve um sistema construído sobre um motor de pesquisa que consegue indexar toda a informação que o utilizador viu ao longo de um determinado tempo, o *Stuff I've Seen* [Dumais03]. Esta informação pode ser sobre páginas na Web, emails e documentos, entre outros. Os resultados dessas pesquisas são apresentados com a ajuda de uma *timeline*, em forma de vista geral ou em detalhe. A vista sumária permite ver a distribuição dos resultados da pesquisa ao longo do tempo. A vista detalhada permite analisar com mais detalhe um determinado resultado da pesquisa. A *timeline é anotada* por marcos pessoais (fotos, tarefas) e públicos (noticias, férias). Este sistema limita-se a efectuar pesquisas simples sem uma análise semântica da informação..

Todos os trabalhos referidos permitem a visualização de informação pessoal. No entanto descuram alguns aspectos importantes, considerados pelo PersonalNews. Por exemplo, apresentamos a informação de forma a permitir a percepção imediata de padrões semanticamente relevantes na informação. Adicionalmente, a informação é mostrada de forma diferente de acordo com a sua importância relativa, tornando-a evidente aos olhos do utilizador, que pode assim guiar as suas pesquisas.

3. INTERFACE

Na Figura 1 pode ser vista uma imagem da interface do sistema por nós desenvolvido. A interface tem a forma de uma página web, com o aspecto tradicional de um jornal. A aplicação, embora local, corre num navegador comum. O uso de HTML e CSS dá-nos a versatilidade de incluir todo o tipo de texto e mesmo imagens nas notícias.

Embora a interface ainda se encontre na fase inicial do seu desenvolvimento, é possível identificar todos os elementos relevantes da mesma. No topo, encontramos o título do jornal, seguido da data a que se reporta (Figura 2). Em seguida, dispostas de acordo com a sua importância relativa (inferida a partir do número de documentos de cada tema), surgem as várias notícias. Cada tema pode resultar em uma ou mais notícias, de diferentes relevâncias. As notícias com mais destaque são mostradas, tal como nos jornais reais, com um título maior e ocupando uma parte maior do ecrã.

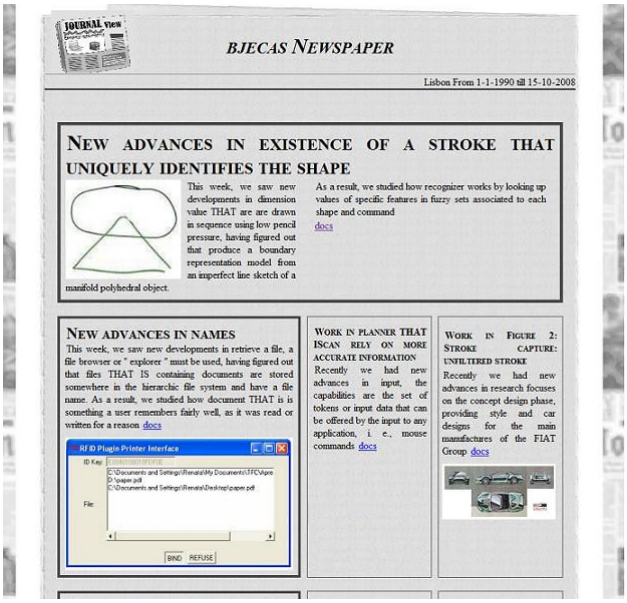

Figura 1 –PersonalNews (Protótipo)

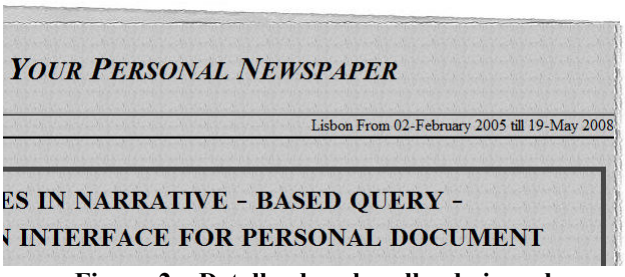

Figura 2 – Detalhe do cabeçalho do jornal

A disposição das notícias na página é decidida automaticamente com base na quantidade de notícias a apresentar e a importância destas. É usado um algoritmo de sub-divisão sucessiva do espaço para o fazer. A página começa por ser dividida em duas, ficando logo uma das partes reservada para a notícia de maior importância. De seguida, divide-se a parte ainda não reservada em mais

duas, sendo novamente uma das partes reservada para a segunda notícia mais relevante, e assim sucessivamente até todas as notícias terem sido posicionadas. O tamanho reservado para cada uma depende da sua importância relativamente às restantes. Ao dar mais destaque a temas mais importantes, fornece-se imediatamente uma pista visual ao utilizador sobre as suas actividades num determinado período de tempo.

Para navegar entre várias edições do jornal pessoal para seleccionar intervalos de tempo diferentes a que o jornal se reporta carrega-se sobre a data (Figura 2) o que nos leva uma secção da interface onde se podem efectuar as operações referidas. Esta contém uma barra temporal onde se encontram as edições do jornal já visitadas, sob a forma de sub-barras nos respectivos intervalos de tempo (Figura 3). Consegue-se assim ter uma percepção geral de todas as edições que existem. Para se criar uma edição do jornal referente a um período novo, basta seleccionar um novo intervalo de tempo em menus de selecção da data.

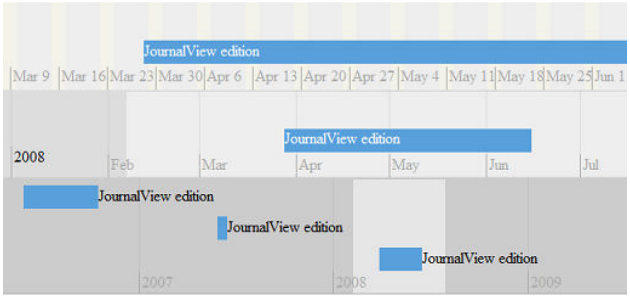

Figura 3 – Barra de temporal para selecção de edições

4. DESCOBERTA DE TEMAS

Qualquer utilizador trabalha em várias áreas e temas ao longo do tempo. Isto reflecte-se nos conteúdos dos documentos por ele produzidos ou consultados. A indexação desses documentos e a subsequente análise desse índice possibilita a organização destes em vários temas, representativos das áreas de interesse do utilizador, e que estão na base das várias noticias criadas.

A descoberta de temas é feita através do agrupamento de documentos através de *clustering*. Numa primeira fase, são extraídas as palavras existentes em cada um dos documentos, bem como a sua frequência dentro desses documentos. As mais relevantes são escolhidas como sendo as palavras-chave representativas do documento. Essas palavras-chave são a base para o agrupamento de documentos por temas. Na sua posse, aplica-se a *Latent Semantic Analysis* (LSA) para encontrar as que melhor caracterizam os vários temas subjacentes aos documentos, estabelecendo uma medida de distância entre eles. Com esse resultado o algoritmo de *clustering* QT-Clust cria grupos de documentos do mesmo tema.

4.1 Escolha das Palavras-Chave

O método usado para a extracção das palavras-chave foi o TFIDF [Brooks06]. Informalmente, esse algoritmo escolhe para um documento palavras que ocorrem frequentemente neste e raramente nos demais. São assim escolhidas as palavras-chave que melhor caracterizam os documentos, e não as mais frequentes, o que iria resultar na captura de, por exemplo, artigos e verbos comuns, que não discriminam na realidade entre os documentos.

4.2 QT-Clust

Após a análise dos vários algoritmos de *clustering* existentes, concluimos que um algoritmo adequado para o nosso problema é o QT-Clust [Heyer99]. O algoritmo funciona da seguinte maneira: um grupo candidato é criado com base no primeiro elemento dos dados e junta-se a este um outro elemento que esteja próximo. Vão-se juntando outros elementos a este grupo candidato iterativamente, sendo que a distância do grupo a cada elemento que se junta é menor que um certo raio pré-determinado. Quando este raio é ultrapassado, é criado um novo grupo candidato, começando com o segundo elemento dos dados repetindo-se o procedimento, tendo novamente em conta todos os elementos que formam o conjunto de dados. No fim deste processo, têm-se tantos grupos candidatos como elementos do conjunto de dados. O maior é guardado, retirando-se os elementos que fazem parte deste do conjunto de dados e repete-se o processo todo novamente. Escolheu-se este algoritmo devido ao facto de, para além de garantir a criação de grupos com qualidade, não é necessário saber à partida o número de grupos do conjunto de dados. Esta é claramente uma vantagem dado ser impossível prever esse número, no nosso domínio de aplicação. A desvantagem da sua utilização consiste na necessidade de fornecer um raio a partida, raio este que vai ter influência na altura da criação dos grupos, por marcar a distância máxima de separação entre elementos do mesmo grupo. No entanto, com a normalização das distâncias entre os elementos, este problema é atenuado, podendo-se estimar um raio adequado para o caso geral, como descrito na secção 4.6..

4.3 Matriz de Distância

Como se pode constatar, para que o algoritmo possa ser aplicado é necessária a existência do conceito de *distância* entre dois elementos (no nosso caso, documentos). Assim, é criada uma matriz de distâncias entre documentos. Nesta matriz, as linhas contém os documentos e as colunas contém as palavras-chave existentes nos documentos. Cada posição da matriz indica quão bem é um documento caracterizado por uma palavra-chave em particular. As linhas da matriz podem ser encaradas como coordenadas num espaço n-dimensional, sendo as distâncias entre os documentos calculadas pela distância euclideana essas coordenadas.

Como iremos agrupar documentos de um intervalo temporal bem definido e limitado, a matriz não será demasiado grande. No entanto, o facto de uma palavra ocorrer num documento não quer necessariamente dizer que seja um bom descritor do tema deste. Reciprocamente, palavras que não ocorrem num documento (embora o façam em documentos relacionados) podem ser determinantes para a descrição do seu tema. Assim, registar nesta matriz apenas a frequência com que as palavras ocorrem nos documentos pode não ser a melhor escolha. A aplicação do algoritmo LSA, descrito abaixo, permite em simultâneo reduzir a

dimensão da matriz e encontrar relações à priori escondidas entre certas palavras e documentos.

4.4 LSA

De uma forma geral, o algoritmo LSA [Berry95] [Landauer94] permite descobrir relações ocultas entre palavras-chave e documentos. Com base numa matriz documento/palavra-chave como a descrita acima, em que cada posição representa o número de vezes que uma palavra aparece num documento, é aplicada uma SVD (*Simple Value Decomposition*). A SVD divide a matriz documento-palavra-chave, A, em três matrizes: U, S e V, tal que $A = USV^T$. A matriz U diz respeito aos termos, isto é, é a matriz cujas colunas são os vectores próprios da matriz AA^T, sendo as coordenadas dos vectores de termos individuais. A matriz S contém os valores mais representativos da matriz documento-palavra-chave, ou seja, é a matriz cujos elementos da diagonal são os valores próprios da matriz A. A matriz V contém os documentos, é a matriz cujas colunas são os vectores próprios da matriz A^TA, representado assim as coordenas dos vectores de documentos individuais. De seguida escolhe-se um valor k, tanto maior quanto mais se quiserem simplificar as matrizes, e substituem-se os valores das k ultimas colunas destas por zero. Efectivamente, uma das propriedades da SVD é colocar os termos com menor importância no fundo da matriz. Ao trocar os valores correspondentes por zero, estão a eliminar-se de consideração termos que, apesar de identificados pelo TF-IDF, têm na realidade pouco poder descritivo face aos temas em que os documentos se inserem. As matrizes resultantes desta substituição, U', S' e V', podem então ser multiplicadas para reconstituir a matriz original, com duas diferenças importantes: apenas são tidas em conta as palavras-chave mais descritivas, e palavras-chave que não ocorrem em determinados documentos podem mesmo ser associadas a estes, caracterizando-os, por aparecerem em documentos relacionados. Em suma, obtem-se uma melhor descrição dos documentos face aos seus temas.

4.5 Junção de Grupos

Embora os resultados obtidos através do algoritmo de *clustering* geralmente fossem completos, notava-se uma clara tendência para a produção de demasiados grupos. Isto é preferível ao agrupamento de documentos díspares no mesmo grupo, mas fragmenta os temas, fazendo com que documentos que deveriam estar juntos estejam em grupos diferentes. Assim, foi necessário criar mais um algoritmo que junte grupos segundo o seguinte critério: determinam-se os centro dos grupos, verificando a distância entre estes. Se esta distância for menor do que um determinado valor, juntam-se os grupos. Após aplicar o algoritmo a conjuntos de teste, verificou-se que se obtem um número de grupos perto do esperado para esses conjuntos, cujos temas eram conhecidos à priori.

4.6 Medidas de Avaliação e Testes

O processo de agrupamento que acabámos de descrever depende do valor de alguns parâmetros: o número de palavras-chave utilizadas, o raio de agupamento e a distância de junção dos grupos. Para saber qual o valor mais adequado desses parâmetros na aplicação em causa, foi necessário efectuar vários testes, variando os seus valores e comparando os resultados.

Usaram-se duas medidas principais de comparação: *Silhouette Coefficient* [Kaufman90] e a Taxa de Sucesso de Agrupamento. O primeiro mede a coesão dos grupos, e o segundo a qualidade do agrupamento. O *Silhouette Coefficient* é uma medida própria dos algoritmos de *clustering*, que mede a qualidade dos grupos, independente do seu número. Dado um ponto i num grupo A, então a(i) é a distancia média entre o ponto i e os restantes pontos contidos em A, e b(i) é a distancia média entre o ponto i e os pontos dum cluster B mais próximo. O *Silhouette Coefficient* de de i é dado por $s_i = (b_i - a_i)/\max(a_i,b_i)$. O *Silhouette Coefficient* de um grupo é dado pela média de todos os dos seus elementos. Se $0.7 \leq SC(P) \leq 1.0 \rightarrow$ separação excelente entre os clusters; $0.5 \leq SC(P) < 0.7 \rightarrow$ separação média entre os clusters (os grupos possuem um centro bem definido); $0.25 \leq SC(P) < 0.5 \rightarrow$ separação fraca entre clusters.

Em relação à Taxa de Sucesso, e considerando que termos os documentos dos conjuntos de teste previamente classificados, é o número de documentos colocados no grupo certo. Uma vez normalizado, o valore, idealmente, o deveria ser próximo de um.

Para os testes foram usados dois conjuntos de documentos pré-classificados em temas. Para estimar o número de palavras-chave e o raio de agrupamento mais adequados, fizemos variar esses mesmos valores em simultâneo, uma vez que ambos os aspectos estão relacionados. Verifica-se (Figura 4) que o número de palavras-chave ideal é de cerca de 30, tendo-se confirmado que o valor do raio deverá ser 0.22. Com estes valores obtêm-se bons resultados no *Silhouette Coefficient*, com um valor de 0.6, o que significa que conseguimos separação média mas clara entre os grupos.

Com a análise da taxa de acerto (figura 3) conseguimos comprovar os novos valores. Para o teste a que a figura se refere, variámos os raio entre 0.1 e 0.3 o número de palavras-chave entre 10 e 100 (a região com melhores resultados) e determinamos para cada um a taxa de acerto. Consegue-se verificar que para um raio próximo de 0.22 e com 30 palavras-chave a taxa de acerto está acima de 50%.

Embora estes valores não sejam os que, em termos absolutos, produzem os melhores resultados, são valores onde os resultados são bons para ambos os conjuntos de dados, sendo um bom compromisso. Efectivamente, para a nossa aplicação não são necessários resultados perfeitos neste nível, uma vez que a extracção de sumários depura e melhora o resultado final.

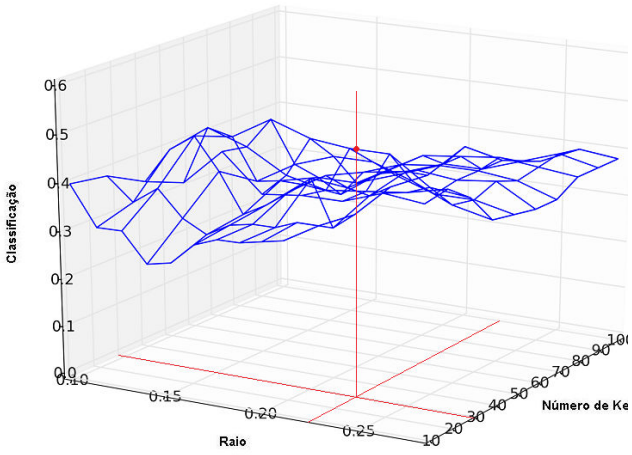

Figura 4 – Classificação obtida com para os diferentes raios e diferentes números de palavras-chave.

Em relação à distância de junção de grupos, chegámos a conclusão que o melhor valor para este seria de 0.2. Este valor faz sentido, dado que está perto do raio usado no algoritmo de *clustering*. Estamos, portanto, a juntar grupos que o proprio algoritmo quase juntou, a menos de artefactos do seu funcionamento. Mais uma vez, o valor do *Silhouette Coefficient* continuava bom, sendo para este caso ligeiramente melhor, embora este melhoramente fosse apenas na ordem das milésimas, sendo este 0.61. Previsivelmente, com distâncias de junção maiores, este valor aumentava ligeiramente, mas à custa da junção indevida de muitos grupos (menor taxa de sucesso).

5. GERAÇÃO DAS NOTICIAS

Na posse dos grupos de documentos, importa conseguir uma descrição dos temas de cada um. Esta informação deve descrever o grupo para que, em seguida, possam ser criadas notícias sobre cada tema.

5.1 Extracção de Palavras-Chave dos Grupos

Para conseguir extrair a informação mais relevante dos vários documentos num grupo, é necessário encontrar as palavras-chave que melhor os caracterizam. Assim, usa-se novamente o algoritmo TF-IDF para a extracção das palavras-chave dos documentos que estavam contidos em cada grupo. Estas palavras-chave, ao usar apenas os documentos de um grupo, são mais informativas e descritivas do que as originais, melhor caracterizando o grupo na sua totalidade. Foi calculada a frequência com que aparecem nos vários documentos e dada prioridade às mais frequentes na extracção dos sumários.

5.2 Extracção dos Sumários

As palavras-chave caracterizam correctamente os grupos. No entanto, para a geração de notícias, as palavras-chave não bastam. Efectivamente, as palavras, isoladamente, não possuem uma semântica suficiente para descrever de forma articulada os temas a que se reportam. Assim, estas foram usadas como ponto de partida para a obtenção de sumários. Para o fazer percorrem-se os documentos de um grupo à procura de frases/orações onde ocorram palavras-chave. Uma vez encontrada uma palavra-chave, procuram-se outras dentro de um raio pré-determinado. O processo é recursivo: se for encontrada uma outra palavra-chave dentro desse raio, procuram-se outras tendo essa como base. Isto identifica um excerto de um documento. Se o número de palavras-chave nesse excerto for maior que um número pré-estabelecido estamos na presença de um sumário/oração relevante para o grupo. Um mínimo de duas palavras mostrou-se razoável, visto não fazer sentido considerar todas as frases onde aparece uma única palavra-chave, e ao considerar um número maior que dois corríamos o risco de obter poucas frases. Ao efectuar este processo para todos os documentos do grupo, ficamos na posse de um vasto leque de informação semanticamente rica acerca do tema desse grupo.

Após a extracção dos sumários para cada grupo é necessário tratar estas orações/frases para que estas apenas possuam informação relevante, e não necessariamente frases completas. Queremos também evitar situações em que o raio usado cause a cisão indevida de orações com relevância.

Um dos primeiros tratamentos efectuados à frase/oração, é a simplificação desta através das vírgulas. Como a maioria das frases/orações possuem vírgulas para separar as várias partes da frase, retem-se apenas o texto até essas vírgulas (ou a partir delas). No entanto, houve o cuidado de deixar de ter em conta as vírgulas quando se tratava de uma enumeração, como por exemplo a seguinte frase: *[…] we considered the follow objects: spheres, squares, circles, triangles […]*. Um raciocinio semelhante foi usado usando como fronteira as preposições como *that* e *which* indicam uma clara divisão na frase.

5.3 Análise Léxical

Não é possível usar directamente os sumários para a criação das notícias. É necessário saber, por exemplo, em que tempo verbal se encontra a oração, se está no singular ou plural, etc, para que os textos criados façam sentido. Assim sendo foi necessário incluir nesta parte do trabalho a análise lexical, âmbito de Língua Natural, e outros métodos desenvolvidos para encontrar o tempo verbal de uma determinada frase.

A análise lexical propriamente dita foi realizada com a ajuda do NLTK, um *toolkit* de língua natural em Python. Neste criaram-se uma série de expressões regulares para identificar a morfologia das palavras, isto é, se uma determinada palavra é um nome, verbo, adjectivo, etc. No nosso caso apenas interessavam os nomes, para saber se uma determinada oração estava no plural ou no singular, e os verbos, para conhecer o seu tempo verbal. Como só estamos a considerar documentos escritos em inglês, não foi muito difícil efectuar esta análise.

5.4 Padrões de Notícias

Para organizar os sumários obtidos num todo coerente, recorremos a padrões (*templates*) de possíveis notícias, em que esses sumários são usados de forma adequada. Como os sumários extraídos poderiam ser orações sob as mais diversas formas, isto é, poderiam começar tanto com um nome ou com um verbo, é necessário criar diferentes

padrões de frases para que, depois de se juntar o texto contido no template e a respectiva oração, o texto fizesse sentido. Visto não se saberem à priori, os temas a reportar, o texto dos padrões é o mais abstracto possível.

Cada padrão possui obrigatoriamente um título e um corpo da notícia. Para além destes, existem elementos opcionais, usados apenas se existirem sumários relevantes para os preencher. É o caso de imagens e excertos de documentos. Finalmente, é possível especificar elementos alternativos, a escolher aleatoriamente durante a criação da notícia para enrriquecer o jornal, mas também para serem usadas em situações específicas. Por exemplo, uma palavra poderá ser usada se o sumário seguinte estiver no passado, e outra se este estiver no presente. Os padrões são descritos em XML, por exemplo:

```
<newsTemplate importance ="1">
    <title>
      <alt>
        <li>New advances in</li>
        <li time="Present">Developments on</li>
      </alt>
      <textExcerpt morphology="name"> </textExcerpt>
    </title>
    <body>
      This week, we saw new developments in
      <textExcerpt quoteSize="small"> </textExcerpt>,
      having figured out that
      <textExcerpt morphology="name"> </textExcerpt>.
      As a result, we
      <alt>
        <li time="present">studied how</li>
        <li time="future">will study how</li>
        <li time="past">had studied how</li>
      </alt>
      <textExcerpt quoteSize="medium"> </textExcerpt>
    </body>
</newsTemplate importance ="1">
```

A utilização destes padrões permite uma fácil adaptação e enriquecimento da interface. Um exemplo de uma notícia (curta) criada automaticamente pelo sistema é:

> **Work in mesh quality and mesh accuracy**
>
> *Recently we had new advances in partitioning techniques that are the most popular methods for rendering implicit surfaces, by creating a polygonal mesh.*

6. CONCLUSÕES

Um dos problemas encontrados hoje em dia ao gerir a grande quantidade de informação pessoal ao nosso dispor no computador é conseguir visualizá-la de forma suficientemente resumida e agrupada de forma a dar ao utilizador uma percepção geral desta e de padrões acerca das suas actividades e interesses ao longo do tempo. A nossa solução, o PersonalNews, permite obter essa percepção ao agrupar os documentos do utilizador de acordo com os seus temas e apresentando esses temas sob a forma de notícias num "jornal pessoal". Conseguimos identificar correctamente os temas em causa, bem como obter excertos relevantes dos documentos que caracterizem esses temas. Esses excertos são então usados para a criação das notícias propriamente ditas.

Testes preliminares com utilizadores mostram que, apesar de no geral bem a nossa abordagem tenha tido sucesso, existem pontos onde se pode melhorar significativamente. Em particular, um dos pontos mais fracos é a qualidade das frases. Os resultados obtidos justificam-se com a não utilização de uma aplicação de língua natural já testada e em funcionamento, sendo aplicada uma criada por nós. No futuro, será interessante estudar a utilização de um sistema de compreensão de língua natural mais maduro, e também o considerar de outras fontes de informação pessoal, como o email ou agenda do utilizador, na criação das várias edições do jornal.

7. REFERÊNCIAS

[Berry95] Berry, M., W., Dumais S.,T. 1995. Using linear algebra for intelligent Information retrieval. SIAM.

[Brooks06] Brooks, C., H. and Montanez, N. 2006. An Analysis of the Effectiveness of Tagging in Blogs Proc. 2006 AAAI Conf.

[Dumais03] Dumais, S., Cutrell, E., Cadriz, JJ., Jancke, G., Sarin, R., Robbins, C. D. 2003. Stuff I've seen: A System for personal information retrieval and re-use. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.

[Kaufman90] Kaufman, L., Rousseeuw, P., J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series In Probability and Statistics. John Wiley and Sons.

[Lamming94] Lamming, M., Flynn, M. 1994. "Forget-me-not" Intimate computing in Support of Human Memory. In Proceedings of FRIEND21

[Landauer94] Landauer, T.,K., Dumais, S., T. 1994. Latent Semantic analysis and the measurement of knowledge. In R. M. Kaplan and J. C. Burstein (EDS) Education Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education.

[Heyer99] Heyer, L., J., Kruglyak, S., Yooseph, S. 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res

[Ringel03] Ringel, M., Cutrell, E., Dumais, S. and Horvitz, E. 2003. Milestones in time: The value of landmarks in retrieving information from personal stores. To appear in the Proceedings of Interact 2003

[Smith06] Smith, G., Czerwinski, M., Mayers, B., Robbins, D., Robertson, G., Tan, D. S. 2006. FacetMap: A Scalable Search and Browse Visualization. In IEEE Transactions on visualization and computer graphics vol. 12

[Viégas06] Viégas, F., Golder, S., Donath, J. 2006. Visualizing Email Content: Portraying Relationships form Conversational Histories. In ACM CHI 2006