Interfaces Baseadas em Narrativas para Recuperação de Documentos Pessoais

Daniel Gonçalves Tiago Guerreiro Joaquim A. Jorge
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
{daniel.goncalves, tjvg, jaj}@inesc-id.pt

Resumo

Um problema frequentemente encontrado por utilizadores de computadores consiste em localizar um determinado documento electrónico. Com efeito, as formas tradicionais de os organizar são cada vez menos eficazes para lidar com a quantidade de informação que mesmo os utilizadores comuns têm ao seu dispor hoje em dia. Uma abordagem inovadora consiste em usar as interacções passadas dos utilizadores com os seus documentos para estabelecer um contexto. Este, exprimível sob a forma de informação autobiográfica, contém toda uma panóplia de informação sobre os documentos, mais fácil de recordar e mais relevante para os utilizadores do que os habituais indícios fornecidos pelo sistema operativo. Apresentamos um novo paradigma de interacção, Interfaces Baseadas em Narrativas, que permite a recuperação de documentos a partir de histórias que os descrevem. Essas histórias, ao estruturarem a informação nelas contida, permitem exercitar a memória associativa dos utilizadores para mais facilmente dela se recordarem.

Com base na avaliação, por parte de utilizadores, de um protótipo funcional, confirmámos que estes são capazes de contar as suas histórias ao computador como se de um ouvinte humano se tratasse, e que as histórias são suficientemente precisas na descrição dos documentos, contendo informação fidedigna sobre os mesmos. Foi também possível mostrar que as histórias possuem um poder discriminativo suficiente para identificar documentos concretos, e que é possível recuperar simples e eficazmente documentos textuais e não textuais. Estes resultados demonstram a validade da abordagem por nós estudada: Interfaces Narrativas para Recuperação de Documentos pessoais.

Palavras-Chave

Interfaces Baseadas em Narrativas, Recuperação de Documentos Pessoais, Informação Autobiográfica, Gestão de Informação Pessoal, Desenho Centrado no Utilizador, Interfaces Pessoa-Máquina

1. INTRODUÇÃO

Duas tarefas que nenhum utilizador de computadores consegue evitar são organizar e posteriormente encontrar os seus documentos electrónicos. Infelizmente, os mecanismos existentes para as realizar não sofreram alterações significativas nas últimas décadas. Baseiam-se, primordialmente, na classificação de *todos* os documentos numa hierarquia definida pelo utilizador, no sistema de ficheiros, e na atribuição de um nome ao ficheiro que representa o documento dentro dessa hierarquia. Isto causa inúmeros problemas, uma vez que não é invulgar um documento poder ser classificado em mais do que uma categoria, ou não parecer enquadrar-se em nenhuma das existentes. As decisões a que esse tipo de classificação obriga aumentam desnecessariamente a carga cognitiva a que os utilizadores estão sujeitos e dificulta a recuperação dos documentos num momento posterior.

Apesar da recuperação de documentos pessoais parecer inserir-se na área da recuperação de informação, um caso especial da pesquisa de documentos em bibliotecas electrónicas ou na Web, não é esse o caso. Efectivamente, enquanto que a recuperação de documentos tradicional se preocupa com permitir aos utilizadores encontrar novos documentos de um determinado tema ou com um determinado conteúdo, a recuperação de documentos pessoais, pela sua natureza, possui aspectos específicos não abordados pelas estratégias ditas clássicas.

Ao tratar-se de documentos pessoais, o utilizador já os conhece a priori e já interagiu com eles no passado. Assim, dispõem de todo um conjunto de informação, para além do conteúdo do próprio documento, que lhe está associado. Essa informação, de natureza autobiográfica, provém do facto de os utilizadores e os seus documentos partilharem um contexto comum. Este inclui aspectos

importantes para os utilizadores e mais fáceis de recordar do que o documento em si ou o seu conteúdo, como por exemplo a razão pela qual foi criado ou lido. Assim, estratégias de recuperação em que apenas o conteúdo ou localização dos documentos são usados ficam necessariamente aquém da realização do potencial subjacente a toda a informação autobiográfica de que os utilizadores se recordam. Isto é verdade mesmo para os sistemas mais recentes e conhecidos, como o Google Desktop, que foca o seu funcionamento na pesquisa de documentos a partir de palavras-chave ou expressões no seu conteúdo. Apesar de, por motivos históricos, isto ser compreensível (o Google permite pesquisas na Web), limita a pesquisa a documentos textuais, ignorando toda a informação adicional supracitada.

Vários trabalhos descritos na Secção 2 tentaram já fazer uso de subconjuntos de informação autobiográfica, mas sem a considerar na totalidade e sem seguir novos princípios organizadores das suas interfaces para que a utilização da informação fosse tão eficiente e eficaz quanto possível. Falta uma interface que permita ao utilizador, de forma natural, referir toda a informação relevante. No nosso trabalho desenvolvemos um novo paradigma de interacção, interfaces baseadas em narrativas, que o torna possível. Com efeito, os seres humanos são contadores de histórias por natureza [Brown91]. Todos contamos histórias desde a infância até à velhice. Em particular, contamos histórias sobre os documentos que procuramos quando temos alguém ao nosso lado durante essa pesquisa. Uma interface capaz de fazer uso dessas histórias para recuperar os documentos é, ao mesmo tempo, natural para os utilizadores, e faz uso da informação realmente associada pelos utilizadores aos seus documentos. Adicionalmente, ao tratar-se não de uma colecção aleatória de propriedades, mas de um todo coerente e interligado, uma história poderá ser capaz de levar os utilizadores a recordar mais informação útil, fazendo uso da sua memória associativa.

Recorrendo a estudos com utilizadores e com a ajuda de um protótipo funcional foi possível, como veremos neste artigo, demonstrar a validade das Interfaces Baseadas em Narrativas para Recuperação de Documentos Pessoais. Demonstrámos ser possível recuperar documentos textuais e não textuais usando narrativas, incluindo alguns que soluções tradicionais não encontrariam.

Na secção seguinte referiremos algum trabalho relacionado, ao que se seguirá uma descrição sucinta da investigação que conduziu a uma profunda compreensão das histórias sobre documentos e à criação de um protótipo funcional de uma interface para recuperação de documentos baseada em narrativas. Seguir-se-á uma descrição dos vários estudos com utilizadores que nos permitiram verificar que as histórias são, de facto, uma solução para o problema descrito., terminando com a apresentação das principais conclusões da investigação desenvolvida e sugerindo algum trabalho futuro.

2. TRABALHO RELACIONADO

Thomas Malone efectuou um dos primeiros estudos sobre as formas de organizar documentos [Malone83]. Neste, a dificuldade dos utilizadores em classificar todos os documentos ficou bem patente. Era preferível para grande parte dos utilizadores organizar os documentos em pilhas e recorrer à sua memória visual e espacial a ter que os classificar. Hoje em dia o problema subsiste, agravado pelo aumento significativo do número de documentos com que os utilizadores lidam.

A causa da ineficiência encontrada ao usar sistemas de ficheiros hierárquicos consiste nestes não reflectirem a forma mais natural dos utilizadores se referirem aos seus documentos. Quando procuramos um documento, não nos lembramos de um nome ou classificação arbitrariamente atribuídos, mas sim do documento em si, de porque o necessitamos nesse momento, o que continha, etc. Em suma, recorremos a informação autobiográfica quando nos recordamos de um documento. Isto foi confirmado em estudos das caixas de correio electrónico dos utilizadores, em que o mesmo problema de classificação existe ao guardar mensagens nas pastas hierarquicamente organizadas. Verificou-se, no entanto, que muitos utilizadores recorriam a essas mensagens como forma de encontrar documentos! [Whittaker96] Isto deve-se ao facto das mensagens, ao contrário dos documentos no sistema de ficheiros, aparecerem inseridas num contexto que permite mais facilmente encontrá-las, associadas a informação adicional (data e hora, remetente, assunto, etc.) que está mais perto da forma natural de recordar dos utilizadores.

Numa tentativa de permitir aos utilizadores fazer uso dessa informação para recuperar os seus documentos, vários trabalhos foram desenvolvidos. Alguns seguiram uma abordagem mais limitada dando um papel preponderante ao tempo. É o caso do Lifestreams [Freeman96], em que todos os documentos são apresentados sequencialmente, de forma ordenada, sendo possível obter *substreams* de documentos que cumprem um certo critério filtrando a *stream* principal. Já o Timescape [Rekimoto99] apresenta o *desktop* do ambiente de trabalho como uma janela sobre um determinado instante de tempo, que pode ser movido quer para o passado quer para o futuro.

Mais abrangentes, algumas abordagens têm em conta um leque mais alargado de propriedades. O primeiro desses trabalhos foi o Semantic File System [Gifford91], em que os documentos estão organizados em directorios virtuais, resultado de pesquisas segundo critérios como o seu autor ou a data em que foi criado. Mais recentemente, sistemas semelhantes mas mais sofisticados foram desenvolvidos. É o caso dos Placeless Documents de Paul Dourish [Dourish00] e do PACO, de Ricardo Baeza-Yates [Baeza-Yates96], em que os documentos são organizados em colecções que os agrupam por algum critério. Estes sistemas, embora promissores, não resolvem totalmente o problema de recuperar documentos. Recordar valores atribuídos a propriedades arbitrárias não é muito melhor

do que recordar onde, no sistema de ficheiros, está um documento guardado.

Alguns sistemas tentam especificamente fazer uso de propriedades cujos valores não conseguem ser apreendidos directamente dos documentos propriamente ditos mas sim de outras fontes de informação como mensagens de correio electrónico e agendas. É o caso do Haystack [Huynh02] e o Stuff-I've-Seen [Dumais03] que integram informação sobre os documentos recolhida de várias aplicações que o utilizador executa. O MyLifeBits [Gemmell06] vai mais longe, permitindo a integração de informação recolhida por vários dispositivos, tendo como objectivo último o armazenamento de toda a informação relevante para um utilizador, independentemente da sua origem. Apesar de mais completos, estes sistemas requerem que o utilizador comece as suas pesquisas recorrendo a palavras-chave, e só depois pode a informação adicional ser usada para refinar os resultados.

Mais recentemente, surgiram os sistemas de pesquisa no *desktop*, como o Google Desktop, que permitem uma pesquisa, essencialmente, por palavras-chave no conteúdo. Isto fica aquém de aquilio de que os utilizadores se conseguem recordar sobre os documentos.

Alguns sistemas recentes estão baseados na utilização de etiquetas ou *tags*, uma forma de anotação recentemente popularizada. Estes sistemas tornam a organização dos documentos mais simples, ao permitirem que cada documento seja classificado de mais do que uma forma, mediante a atribuição de várias *tags*. No entanto, esta abordagem tem alguns problemas, dada a dificuldade em manter a consistência nas *tags*, o que leva a um baixo nível de reutilização das mesmas. Igualmente, etiquetar todos os documentos pode ser uma tarefa incomportável para a maioria dos utilizadores. O sistema de Ariel Shamir [Shamir04] tenta ajudar o utilizador a organizar os seus documentos mediante *tags*, mas requer que todos os documentos sejam etiquetados manualmente. O Phlat [Cutrell06] também usa *tags* mas em conjunto com outra informação recolhida automaticamente.

3. ESTUDANDO AS NARRATIVAS

Sendo o assunto da nossa investigação a criação de uma interface para a recuperação de documentos pessoais, decidimos envolver os utilizadores em todos os passos da pesquisa realizada. Nos vários estudos descritos abaixo, o procedimento foi similar: procuraram-se utilizadores variados (e não apenas colegas ou amigos) de diferentes formações, profissões e escalões etários. Foram efectuadas deslocações às suas residências ou locais de trabalho e conduzidos os estudos.

Isto tornou o trabalho mais moroso e difícil, ao estar dependente de outros, mas conferiu-lhe uma qualidade impossível de obter de outra forma: só assim poderíamos ter a certeza da adequação dos resultados obtidos às reais necessidades dos utilizadores. Todos os resultados foram estatisticamente analisados com 95% de confiança.

Estudos preliminares permitiram-nos verificar que documentos tem o utilizador ao seu dispor, de que forma estão organizados, etc. Em seguida, avançámos para o estudo das histórias propriamente ditas.

3.1 As Histórias

Dado o objectivo de usar as histórias como forma de permitir aos utilizadores recordarem e referirem informação autobiográfica referente aos seus documentos, importava saber o que delas se pode esperar. Esta era uma área inexplorada, visto que nenhum trabalho se tinha previamente debruçado sobre esta questão. Apesar de existirem trabalhos relacionados com *storytelling* em computadores, estes focam-se prioritariamente na geração de histórias pelo computador, não na sua compreensão, e nunca no domínio específico das histórias sobre documentos. Tivemos, pois, que conduzir um estudo em que, com a ajuda dos utilizadores, pudemos ficar a conhecer detalhadamente as histórias que estes contam sobre os seus documentos, tanto em termos de conteúdo como de estrutura.

Assim **20 utilizadores** foram entrevistados, numa entrevista semi-estruturada, e foram recolhidas **60 histórias** descrevendo vários tipos de documentos reais dos utilizadores [Gonçalves04]. Por um lado, a interacção com um documento criado pelo próprio utilizador é, necessariamente, diferente da com um documento de outro autor. Um documento próprio é conhecido mais intimamente, tendo sido normalmente criado ao longo de um intervalo de tempo maior e requerendo um maior investimento por parte do utilizador. As histórias sobre os documentos do próprio utilizador poderiam, pois, conter informação autobiográfica diferente das histórias sobre documentos de outros autores. Por outro lado, os utilizadores lembrar-se-ão provavelmente de menor quantidade de informação, ou informação menos precisa, se o documento tiver já sido criado há algum tempo. Em busca de eventuais diferenças nas histórias, foram considerados documentos Novos (criados pelo utilizador recentemente), Antigos (criados há pelo menos seis meses) e de Outros (de outros autores).

Os entrevistados tiveram a liberdade de contar as histórias sem restrições. Em seguida, as entrevistas foram transcritas e foram efectuadas manualmente uma análise de conteúdos e uma análise relacional.

Na análise de conteúdos verificámos que as histórias contêm realmente uma **grande variedade de informação** sobre os documentos, que pode ser agrupada nos seguintes elementos (nenhuns outros foram encontrados nas histórias): *tempo, local, autores, co-autores, razão pela qual o documento foi criado, assunto, outros documentos, acontecimentos na vida pessoal do utilizador, acontecimentos no mundo em geral, acontecimentos ocorridos durante a manipulação do documento, tipo do documento, tarefas, local de armazenamento, nome, versões, conteúdo e outras versões do documento.* Toda a informação nas histórias pode ser classificada num destes elementos. Verificámos, também, que as principais diferenças entre as

também, que as principais diferenças entre as histórias se devem ao tipo de documento referido: as histórias sobre documentos do próprio autor são diferentes das de outros. Adicionalmente, as histórias contêm grande quantidade de informação: **as histórias têm em média 17,7 e 12,15 elementos**, para os documentos do utilizador e de outros autores, respectivamente. Ficámos, também, com dados referentes à **frequência relativa dos diversos elementos** nas histórias, sendo possível identificar os mais (tempo, razão de ser, tarefas, conteúdo) e menos (eventos do mundo e durante a manipulação do documento) importantes para os utilizadores. O próprio conteúdo dos elementos também se tornou conhecido, dando-nos uma ideia da informação com q\ue uma eventual interface deveria ser capaz de lidar.

Na análise relacional ficámos a conhecer a **estrutura mais provável** de uma história sobre um documento, a partir das várias probabilidades com que os elementos se podem suceder uns aos outros nas histórias. Com esses dados, foi-nos possível treinar modelos de Markov não observáveis e obter **estruturas de histórias arquetípicas** para os vários tipos de documento.

3.2 Desenhando a Interface

O estudo anterior forneceu-nos uma caracterização completa das histórias sobre documentos, e permitiu a criação de **princípios de desenho de interfaces baseadas em narrativas**. No entanto, esses princípios podiam ser aplicados de diversos modos: a forma que a interface deveria realmente tomar estava ainda por definir. Isto é particularmente importante se atentarmos no facto de estarmos a criar uma interface inovadora num domínio ainda por explorar, não existindo pontos de referência claros para a sua criação.

Assim, foram criados **dois protótipos de baixa fidelidade**, não funcionais, esboçados abaixo. São protótipos criados com cartolina e marcador, que permitem simular o funcionamento de interfaces diferentes sem necessidade de codificações morosas.

Os protótipos criados representam duas abordagens diferentes para a recolha das histórias: no primeiro cada elemento das histórias é representado por um elemento gráfico passível de manipulação directa com o rato; no segundo a história é apresentada textualmente, sob a forma de frases inicialmente incompletas, cujo conteúdo em falta será fornecido pelo utilizador..

Em ambos os protótipos os elementos da história eram introduzidos com a ajuda de diálogos especializados, à esquerda (um para cada elemento), e uma lista de documentos promissores apresentada em baixo, sob a forma de *thumbnails*. Os utilizadores tinham, tal como no caso do estudo anterior, a liberdade para mencionar os elementos das histórias em qualquer ordem. A expressividade de ambos os protótipos, em termos do que é possível mencionar para cada elemento da história, era idêntica: os mesmos diálogos foram usados na avaliação de cada um dos protótipos.

Mais uma vez **20 utilizadores** foram entrevistados e **60 histórias** recolhidas, metade usando cada protótipo. Foram considerados os mesmos tipos de documentos, e as histórias foram registadas para posterior análise estatística. Imediatamente se tornou evidente que **o protótipo em que a história é apresentada textualmente era indubitavelmente melhor**. Não só foram as suas histórias melhores que as do outro protótipo (maiores e com mais informação) como, e mais importante, foram praticamente idênticas às contadas a entrevistadores humanos, tanto em termos de conteúdo como de estrutura.

Concluiu-se assim que o Protótipo B seria a base do trabalho a desenvolver em seguida: a criação de um protótipo funcional, essencial para a continuação da validação das histórias como forma de ajudar a recuperar documentos.

Pudemos também verificar que apesar de ser útil mostrar a história textualmente, pedir aos utilizadores que a introduzissem, na sua totalidade, sob a forma de texto não seria adequado. Isto foi por eles considerado como demasiado moroso e poderia também trazer problemas de interpretação, dado o contexto não restrito das narrativas, em que informação dos mais variados tipos pode ser mencionado pelos utilizadores ao descrever os seus documentos pessoais.

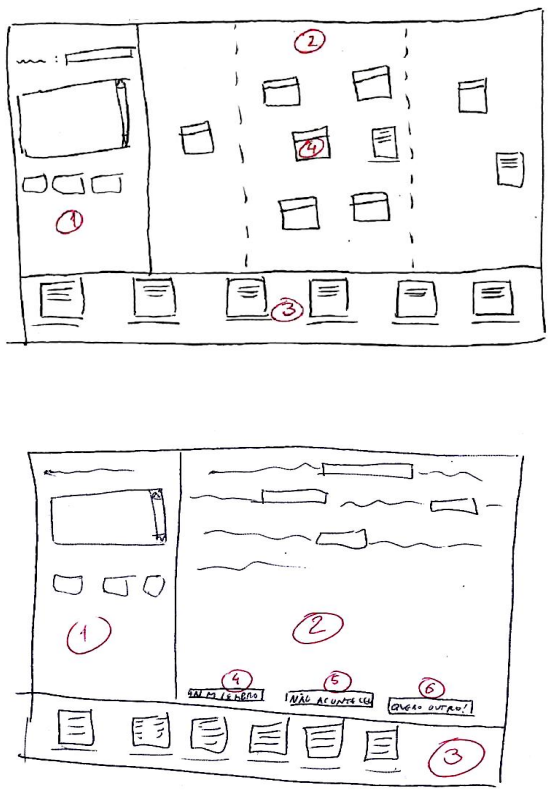

Figura 1–Esboços dos Protótipos de Baixa Fidelidade

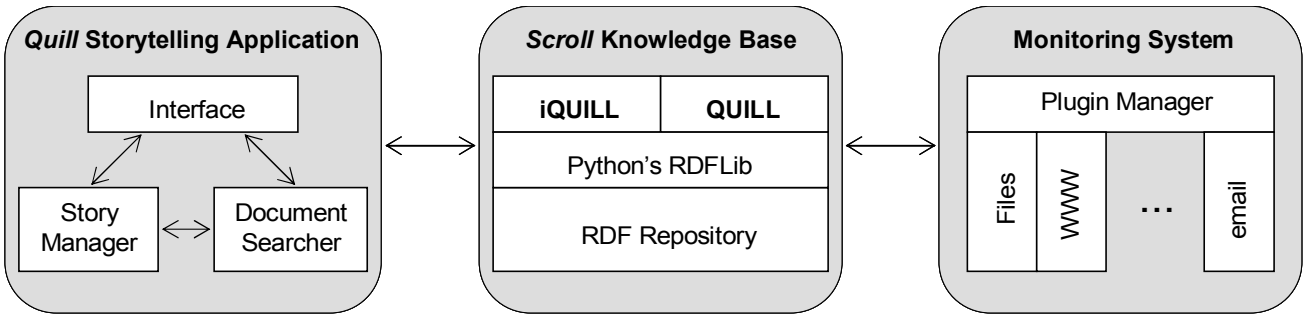

Figura 2 – Arquitectura do Sistema

3.3 O Protótipo

Foi desenvolvido um protótipo funcional da interface escolhida, não como um fim em si mesmo, mas de modo a ter uma ferramenta que permita a elaboração de novos testes tendo em vista uma investigação mais aprofundada do nosso objecto de estudo: as narrativas sobre documentos. A arquitectura geral do sistema criado está esquematizada na Figura 2. Nela, podem destacar-se três partes principais: Uma Base de Conhecimento, um Sistema de Monitorização, e a Interface de Recuperação de Documentos.

3.3.1 A Base de Conhecimento Scroll.

Atentando à descrição das histórias obtida após as entrevistas com os utilizadores, tornou-se evidente que, mais do que manter uma base de dados sobre os documentos dos utilizadores, seria necessária uma base de conhecimento. Isto permitiria ter flexibilidade suficiente não só para representar toda a informação necessária, como para a incorporação de conhecimento do senso comum, sobre o utilizador e o mundo que o rodeia, de modo a melhor compreender as histórias.

Assim, foi feito um levantamento dos principais formalismos de representação do conhecimento actualmente em uso, e decidimos basear a nossa base de conhecimento no formalismo RDF, pela sua simplicidade, escalabilidade, e por se tratar de um standard W3C com níveis de adopção crescentes que poderão vir a tornar disponível uma grande quantidade de conhecimento a médio prazo.

No entanto, o RDF "puro" representa todo o conhecimento sob a forma de triplos `(sujeito, predicado, objecto)`. Isto tornaria morosa a sua utilização, pelo que foi implementada uma biblioteca, denominada *Scroll*, para fornecer capacidades de mais alto nível ao resto do sistema. Assim, é possível criar classes, instâncias, verificar propriedades de objectos, etc. Adicionalmente, pode ser efectuada inferência de nó, de caminho, e foi criado um *schema* RDF, `iQuill`, com a definição de vários *case-frames* correspondentes a conectivas lógicas, permitindo a criação e avaliação de regras de inferência. As inferências foram implementadas de modo a garantir a sua eficiência, e são permitidas inferências parciais, em que podem ser pedidos novos valores válidos para as variáveis à medida das necessidades, em vez de se calcularem todos necessidades, em vez de se calcularem todos *a priori*. Outro *schema*, `Quill`, contém a definição das entidades relevantes para o armazenamento da informação autobiográfica.

3.3.2 O Monitor

O subsistema de monitorização foi criado de modo a recolher informação sobre o utilizador, as suas actividades, e os seus documentos, dados das mais diversas fontes. De facto, nenhum utilizador estaria disposto a fornecer manualmente toda a informação eventualmente relevante para posteriormente ser usada nas histórias. Este subsistema evita essa necessidade.

É baseado em *plugins* referentes a várias fontes de informação. Actualmente, é possível recolher informação sobre: os documentos do utilizador; os emails enviados e recebidos; as páginas Web visitadas; as aplicações usadas; os documentos impressos; a agenda do utilizador. Existem, normalmente, dois *plugins* para cada uma destas fontes. O primeiro recolhe toda a informação existente na altura em que o sistema é pela primeira vez executado, e o seguindo mantém essa informação actualizada ao longo do tempo.

Toda a informação relevante é armazenada na base de conhecimento. Documentos textuais são analisados e deles extraídas palavras-chave, recorrendo ao algoritmo tfidf [Salton88], após terem sido divididos em *tokens* e sujeitos a um algoritmo de *stemming* para obter apenas as raízes das palavras tornando a aplicação mais robusta em termos de tempos verbais, género, número, etc. O algoritmo usado foi o algoritmo de Porter [Porter80]. Dos documentos não textuais é guardada toda a meta-informação neles presente (nos cabeçalhos ID3 de ficheiros MP3, por exemplo). Toda a informação genérica é também recolhida, tal como datas de criação, modificação e acesso, tamanhos, etc. O mesmo se passa para emails, os seus anexos, e as demais fontes de informação.

Apesar de separados, os *plugins* formam um todo coerente. Por exemplo, se o *plugin* que monitoriza os emails enviados encontra um anexo, verifica onde está a informação referente a esse documento, já obtida pelo *plugin* de monitorização de documentos, e guarda a nova informação de que este foi mandado por email a alguém mantendo a base de conhecimento coerente.

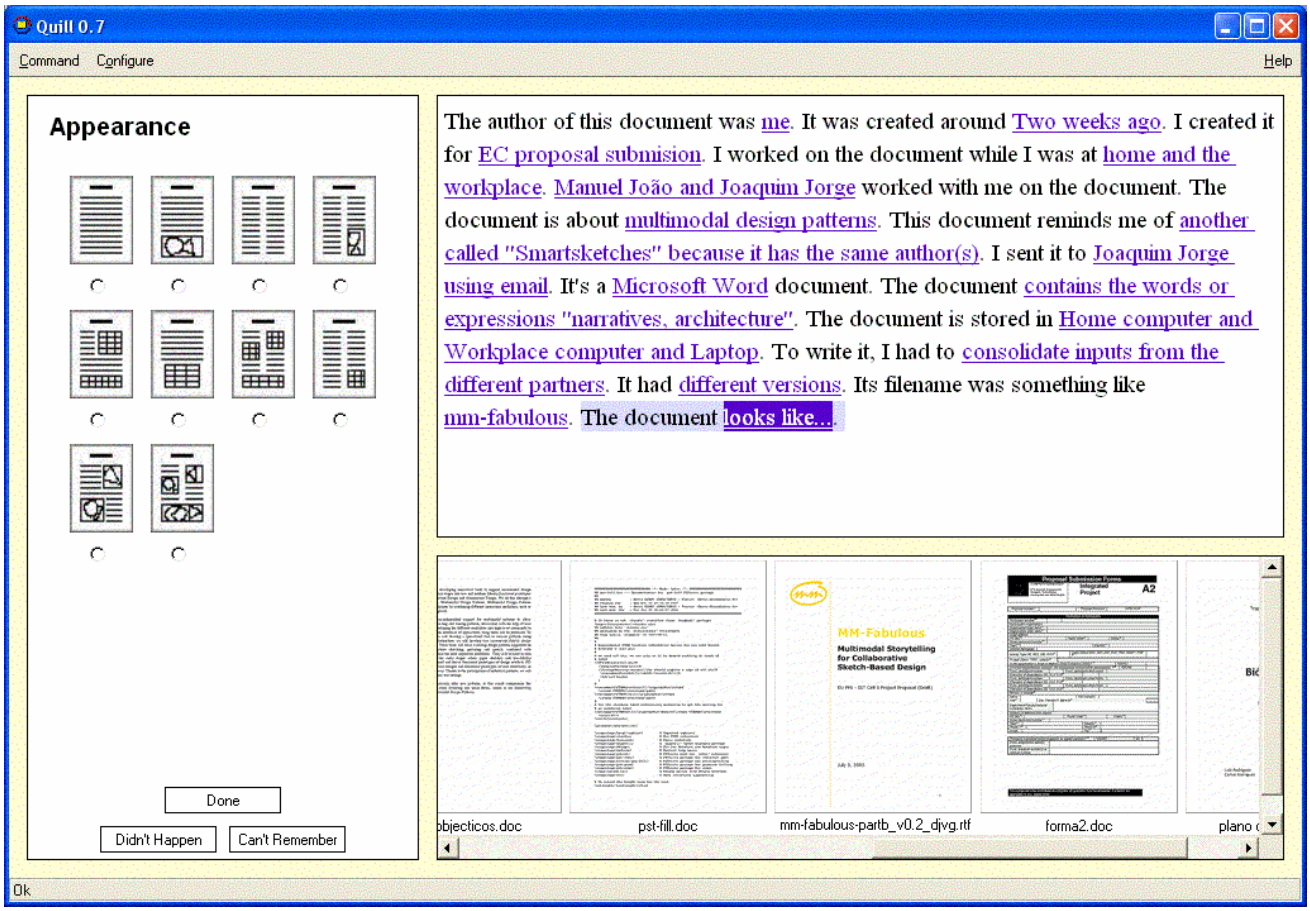

Figura 3 – Aspecto geral da interface

Em termos de implementação este subsistema foi, claramente, o mais complexo e trabalhoso. De facto, dado pretender-se a realização de testes com utilizadores reais e seus documentos, é necessário ter em conta inúmerasvariantes em termos de versões de aplicações e sistema operativo. Por exemplo, foi necessário desenvolver *plugins* para monitorizar email existente nos clientes mais populares: Mozilla Thunderbird, Outlook Express, e Outlook, lidando com várias versões e formas de utilização dos mesmos. Teria sido mais simples escolher apenas um, mas isso inviabilizaria a recolha de dados reais, deitando por terra a fiabilidade dos resultados.

3.3.3 A Interface

A Figura 3 mostra o aspecto da interface Quill, que permite aos utilizadores contarem as suas histórias, e que as utiliza para recuperar documentos. No canto superior direito está a área da história propriamente dita. Nela, as histórias vão sendo incrementalmente construídas. Cada elemento é sucessivamente apresentado ao utilizador sob a forma de uma frase incompleta. Após a introdução da informação, a frase será terminada, sem prejuízo de uma posterior alteração. A informação é introduzida com a ajuda de diálogos especializados à esquerda. A informação pedida em cada um desses diálogos reflecte directamente os conteúdos encontrados nas histórias contadas a contadas a entrevistadores humanos. Os diálogos são flexíveis. Por exemplo, se o utilizador mencionar que um documento é do tipo Microsoft Word, então no diálogo referente ao Conteúdo do documento (em que o aspecto visual do mesmo pode ser especificado) apenas serão dadas opções referentes a possíveis aspectos de documentos desse tipo. A história é assim construída como um todo coerente e legível pelos utilizadores.

Não é possível escrever toda a história na forma de texto livre. Isso seria demasiado moroso para os utilizadores (muitos manifestaram o seu desagrado quando confrontados com essa possibilidade) e traria problemas adicionais à sua compreensão. No entanto, em alguns diálogos, é feito *parsing* de texto, se necessário, primeiro recorrendo a um *chart parser* e a uma gramática livre do contexto aumentada com fórmulas em cálculo lambda, capaz de gerar automaticamente a semântica das frases interpretadas. Se esse mecanismo falhar, efectua-se um *chunk parsing* que, embora menos poderoso, é mais tolerante a erros e permite, mesmo assim, extrair algum significado das frases introduzidas pelos utilizadores.

A ordem na qual os vários elementos da história são sugeridos pelo programa é a inferida a partir das histórias contadas a humanos, que considerámos ser a mais natural. Apesar disso, o utilizador tem ao seu dispor um conjunto de botões que lhe permitem alterar o fluxo da história,

podendo escolher referir outro elemento em qualquer altura, ou afirmando que algo não aconteceu ou que não se lembra de determinada informação. A ordem não é rígida, sendo gerada a cada momento a partir de um modelo de Markov. Este é ajustado cada vez que uma história é contada, de modo a que a ordem sugerida possa reflectir as particularidades de cada utilizador em utilizações subsequentes.

Em baixo, encontra-se a área de sugestão de documentos. À medida que a história se vai tornando mais completa, o Quill procura continuamente os documentos que mais se adequam à descrição nela contida. O seu nome é mostrado junto a um pequeno *thumbnail* da primeira página do documento (ou ao seu aspecto geral, se for uma imagem, por exemplo). Um dos factores mais vezes referidos nas histórias era o aspecto visual dos documentos, pelo que apresentar este *thumbnail*, será útil, tirando partido da memória visual dos utilizadores e ajudando-os a reconhecer o documento procurado, sem perturbar indevidamente o processo de contar a história. Este aspecto mostrou-se tão relevante, que foi separado dos restantes elementos, dando origem a um novo: a *aparência de um documento.*

Sempre que é contada uma nova história, são criadas regras de inferência criadas a cada momento de modo a reflectir a informação introduzida, e reportando-se ao conhecimento recolhido pelo subsistema de monitorização e a conhecimento de senso comum. As inferências correspondentes são então efectuadas no Scroll, e aos documentos resultantes atribuído um valor numérico. Da adição dos valores resultantes das diversas regras de inferência resulta uma pontuação global para os documentos. Os mais bem pontuados serão sugeridos ao utilizador como potenciais soluções.

Um algoritmo por nós desenvolvido garante a sugestão dos elementos melhor pontuados sem requerer, em grande parte dos casos, que todas as regras de inferência sejam avaliadas até ao fim.

4. AVALIAÇÃO

Para atingir o nosso objectivo, demonstrar a validade das narrativas como princípio organizador da informação para recuperação de documentos pessoais, foi necessário responder às seguintes questões de pesquisa:

- É possível contar histórias sobre documentos ao computador sem problemas, como se de um humano se tratasse?
- São as histórias sobre documentos fiáveis, contendo informação verídica sobre estes?
- Têm as histórias um poder discriminativo suficiente para identificar documentos concretos?
- É realmente possível recuperar documentos usando narrativas?

De modo a dar resposta a essas questões, foi necessário avaliar o protótipo com a ajuda dos utilizadores. Esta avaliação foi efectuada em dois momentos distintos. Numa primeira fase, verificou-se se é possível aos utilizadores contarem as suas histórias ao Quill de forma semelhante ao que tinha sido observado com ouvintes humanos. Foi, ainda, verificada a fidedignidade das histórias. O protótipo foi melhorado com base nas lições aprendidas durante esta primeira avaliação. Seguiu-se a segunda avaliação do mesmo, em que foi verificado o poder discriminativo das histórias e quais as taxas de sucesso de recuperação atingidas usando o Quill.

Não foram efectuadas as medidas mais habitualmente encontradas em estudos sobre recuperação de informação devido a estarmos a tratar de documentos pessoais, procurando fazer uso da informação autobiográfica que se lhes refere. Como tal, as técnicas tradicionais de avaliação de mecanismos de recuperação de informação deixam em parte de fazer sentido, para além de serem extremamente difíceis de aplicar. É comum partir de conjuntos de teste de documentos pré-definidos e bem conhecidos, sabendo-se de antemão quais os que satisfariam uma determinada pesquisa. Isto permite a recolha de dados referentes à recuperação que permitem o cálculo da *precision* e *recall*, definidas como a percentagem de documentos relevantes devolvidos pela pesquisa face a todos os devolvidos, e o rácio entre o número de documentos relevantes devolvidos e todos os relevantes existentes no conjunto.

Ao lidar com documentos pessoais, não é possível criar a priori conjuntos de documentos conhecidos. Qualquer avaliação correcta e isenta terá que incidir sobre os documentos dos utilizadores uma vez que apenas sobre estes existe informação autobiográfica. Tentar gerar conjuntos de teste com base nesses documentos focaria a atenção do utilizador sobre os mesmos, invalidando o estudo. Consequentemente, não é possível medir a *recall*, dado que nunca saberemos que documentos existem e quantos seriam relevantes. Quanto à *precision*, não faz sentido medi-la, porque o utilizador não busca *algo* que satisfaça os seus critérios, mas sim *um documento específico* (ou um conjunto limitado de documentos conhecidos). A *precision* corresponderá, efectivamente, ao sucesso da pesquisa. Constata-se, assim, que a recuperação de documentos pessoais usando informação autobiográfica é um problema diferente de grande parte do que já foi estudado anteriormente, sendo necessário o desenvolvimento de novas técnicas e abordagens criadas especificamente para esse domínio.

4.1 Qualidade das Histórias e Fidedignidade

Para realizar este estudo, deslocámo-nos às casas e locais de trabalho de 10 utilizadores e recolhemos 30 histórias. Essas histórias foram contadas usando o protótipo funcional, após a indexação dos documentos reais dos utilizadores.

Avaliámos a qualidade das histórias comparando as histórias contadas ao protótipo às contadas anteriormente na presença de humanos. Se a interface fosse capaz de as

compreender correctamente, deveriam ser semelhantes. Verificámos ser esse o caso. Tanto em termos de estrutura como de conteúdo as histórias são semelhantes. O seu comprimento é essencialmente idêntico, bem como a ordem em que os elementos são mencionados: desvios da ordem, considerada mais provável nas histórias contadas por humanos, ocorreram apenas **0.1** vezes por história. Quanto à importância relativa dos elementos, foi possível constatar que a única diferença significativa se deu para o elemento "Nome", referido **39%** mais vezes do que nas histórias anteriores.

Outro aspecto também por explorar era a fidedignidade da informação autobiográfica referida pelos utilizadores nas suas histórias. Efectivamente, apesar de vários sistemas terem feito uso de parte desta informação, não tinha ainda havido uma preocupação com a veracidade da mesma, assumindo-se que esta estaria correcta. Consideramos isto um aspecto fulcral de qualquer abordagem desse tipo, visto que informação incorrecta poderia comprometer seriamente o funcionamento do sistema. Decidimos pois, com a ajuda do protótipo funcional, realizar mais um estudo com utilizadores em que foi verificado se a informação contida nas histórias sobre documentos, apesar de bastante, é ou não verdadeira.

A fidedignidade de cada uma das 30 histórias recolhidas foi confirmada: após o término de cada história, cada um dos seus elementos foi verificado numa entrevista com o utilizador. Alguns elementos eram passíveis de uma verificação inequívoca (nomes de ficheiros, por exemplo). Outros, pela sua natureza, não podiam ser verificados dessa forma. Por exemplo, se um utilizador mencionar que imprimiu um documento, a menos que nos mostre o documento em papel não temos forma de verificar se isso é ou não verdade. Assim, considerámos dois graus de certeza: elementos verificados para além de qualquer dúvida, e aqueles que após uma investigação cuidada parecem correctos, mas sob os quais subsistem dúvidas.

Com base nos critérios que acabámos de referir, e após uma análise estatística, verificámos que **entre 81% e 91% do que os utilizadores mencionam nas histórias corresponde à verdade**. Ou seja, podemos esperar que entre um e três elementos numa história sejam errados. Isto está dentro de limites perfeitamente aceitáveis, uma vez que a informação de cada elemento é considerada não de forma taxativa mas apenas como mais um indício sobre o documento.. Foi também verificado que, apesar de por vezes errada, a informação não anda longe da verdade (palavras trocadas no nome de ficheiros, por exemplo), pelo que é possível lidar, até certo ponto, com as incorrecções.

4.2 Poder Discriminativo das Histórias e Sucesso de Recuperação

O Quill foi instalado nas máquinas de 20 utilizadores. Num primeiro contacto, foi-lhes explicado o objectivo do estudo, e iniciada a indexação dos documentos do utilizador. Quando se tornava previsível que esta indexação levaria muito tempo, a entrevista prosseguia no dia seguinte, após o seu término. Isto foi feito para evitar ocupar desnecessariamente o tempo dos utilizadores. Mesmo assim, cada entrevista demorou, em média, mais de uma hora, no total.

A segunda parte da entrevista decorreu da seguinte forma: foi pedido aos utilizadores que procurassem, usando o Quill três documentos diferentes: um Recente, um Antigo e um de Outro Autor (à semelhança do que tinha sido feito em todos os estudos anteriores, para permitir uma comparação directa dos dados recolhidos, se necessário). Analisado o resultado de cada pesquisa, distinguiram-se quatro casos:

- O documento foi encontrado
- O documento foi encontrado "a menos de um *click*" (encontrada a pasta correcta, reconhecida como tal e, após aberta, encontrado o documento, por exemplo)
- O documento não foi encontrado porque o utilizador cometeu algum erro na história
- O documento não foi encontrado.

Especialmente no último caso, procurou descobrir-se qual o motivo que levou o documento a não ser encontrado

Após o término da cada história, usou-se o KB Analyzer, uma ferramenta por nós desenvolvida para inspecção da base de conhecimento, para encontrar todos os documentos que satisfazem a informação contida na história. A contabilização desses documentos deu-nos uma medida do poder discriminativo das histórias.

Tratando-se de um estudo importante para a validação da nossa abordagem, foram estabelecidos, a priori, critérios de sucesso. Nomeadamente, esperávamos encontrar:

- Um poder discriminativo de **1 em 5:** Por um lado, estudos de psicologia cognitiva mostram que a quantidade de informação que os utilizadores conseguem reter na sua memória de curto prazo é de 7 ± 2. O número de documentos a mostrar deveria, pois, estar dentro deste intervalo. Por outro lado, a interface mostra, a cada momento, cinco documentos promissores. O documento procurado deveria ser um deles.
- Uma taxa de recuperação de documentos textuais de pelo menos **75%**.
- Uma taxa de recuperação de documentos não textuais de pelo menos **50%**.

Esperávamos obter uma taxa de sucesso de 75% e não superior devido às dificuldades enunciadas acima nesta secção. Efectivamente, se um produto industrial acabado poderia produzir melhores resultados, o protótipo ao nos-

nosso dispor poderia diminuir ligeiramente a taxa de sucesso ideal. Para além disso, a inexistência de informação autobiográfica antiga durante a realização dos testes poderia ter uma influência negativa na taxa de sucesso. O valor de 50% para documentos não textuais prende-se com o facto de que, na ausência de informação autobiográfica antiga, e em virtude do conteúdo destes ser mais dificilmente explorável que o dos documentos textuais, a taxa de recuperação esperada vir a sofrer mais.

Após análise dos dados recolhidos, verificou-se que as histórias têm um poder discriminativo de **2.51**. Ou seja, em média, cada história descreve correctamente 2.51 documentos. É também interessante notar que a moda do número de documentos descritos por cada história é **um**, tendo sido esse o valor encontrado para 55% das histórias. O objectivo estabelecido para o poder discriminativo das histórias foi assim atingido em pleno.

Quanto às taxas de sucesso de recuperação, verificámos que **87.9%** de todos os documentos procurados foram encontrados. Para documentos textuais houve sucesso em **95.2%** dos casos, e para documentos não textuais a taxa foi de **68.8%**. Em suma ambos valores ficaram bastante acima dos objectivos previamente fixados. Note-se, também, que **7.84%** dos documentos foram encontrados sem a ajuda de informação textual, quer no seu conteúdo, quer no seu nome ou nome da pasta em que se encontravam. Esses documentos não seriam encontrados por ferramentas de pesquisa actuais baseadas em palavras-chave, como é o caso do Google Desktop.

Tendo sido possível responder com sucesso às quatro questões de pesquisa enunciadas no início da investigação, ficou provada a validade das Interfaces Baseadas em Narrativas para a Recuperação de Documentos.

5. CONCLUSÕES E TRABALHO FUTURO

Um dos problemas que mais aflige os utilizadores de computadores é a dificuldade de recuperar documentos que sabemos existirem algures no nosso computador. As formas tradicionais de os organizar e recuperar tornaram-se cada vez mais ineficazes à medida que o número de documentos cresce. Sistemas recentes de pesquisa permitem procurar documentos recorrendo a palavras-chave. No entanto, ao tratar-se de documentos pessoais, existe todo um conjunto de informação autobiográfica, proveniente de interacções passadas com esses documentos, que poderia ser usada para mais facilmente os recuperar.

A nossa investigação conduziu à criação de um novo paradigma de interacção, Interacções Baseadas em Narrativas, que permite aos utilizadores estruturarem os vários elementos relevantes para a recuperação de documentos e, recorrendo à sua memória associativa, recordar mais facilmente informação útil para esse fim. A criação de um protótipo repetidamente avaliado pelos utilizadores permitiu verificar as qualidades intrínsecas da abordagem, validando-a.

Adicionalmente, foi possível obter uma análise detalhada das histórias sobre documentos, que nos dá também uma descrição dos elementos de informação autobiográfica considerados mais relevantes pelos utilizadores. Mais, a metodologia usada para efectuar a análise de narrativas sobre documentos pode ser aplicável a outros domínios.

A nossa abordagem é, no entanto, mais dependente da língua do utilizador do que outras estratégias de recuperação de documentos. Requer também algum tempo para que a informação autobiográfica continuamente recolhida sobre o utilizador e os seus documentos possa ser usada para os recuperar.

Vários aspectos são passíveis de um estudo mais aprofundado como trabalho futuro. Seria interessante estudar a utilização de outras modalidades de interacção para enriquecer as histórias. Fala poderia ser usada para contar a história ao computador, ou para a ler ao utilizador, evitando assim que este tenha que a ler se a quiser rever. Uma interface caligráfica poderia ser usada pelo utilizador para criar um esboço rápido do aspecto visual de um documento, para ajudar à sua recuperação. Adicionalmente, seria interessante aplicar as Interfaces Baseadas em Narrativas para a recuperação de outros objectos que não documentos, incluindo objectos do mundo real como livros, ou mesmo qualquer objecto pessoal do utilizador, sobre o qual este terá certamente uma história que contar.

6. AGRADECIMENTOS

Este trabalho foi financiado em parte pelo projecto BIRD, FCT POSI/EIA/59022/2004.

7. BIBLIOGRAFIA

[Baeza-Yates96] Baeza-Yates, R., Jones, T. and Rawlins, G. **A New Data Model: Persistent Attribute-Centric Objects**, Technical Report, University of Chile, 1996

[Brown91] Brown, D. E. **Human Universals**. New York: McGraw-Hill 1991.

[Cutrell06] Edward Cutrell, Daniel Robbins, Susan Dumais and Raman Sarin. **Fast, flexible filtering with Phlat**. In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 261–270. ACM Press, New York, NY, USA, 2006b. ISBN 1-59593-372-7.

[Huynh02] David Karger David Huynh and Dennis Quan. **Haystack: A platform for creating, organizing and visualizing information using RDF**. In Semantic Web Workshop, The Eleventh World Wide Web Conference 2002 (WWW2002). 2002.

[Dourish00] Dourish, P. *et al*. **Extending Document Management Systems with User-Specific Active Properties.** *ACM Transactions on Information Syst.*,18(2), pp 140-170,ACM Press 2000.

[Dumais03] Dumais, S. T. et al. **Stuff I´ve Seen: A system for personal information retrieval and re-use.** In *Proceedings of SIGIR 2003*.

[Freeman96] Freeman, E. and Gelernter, D. **Lifestreams: A Storage Model for Personal Data**, *ACM SIGMOD Record,*25(1), pp 80-86, ACM Press 1996

[Gemmell06] Jim Gemmell, Gordon Bell and Roger Lueder. **Mylifebits: a personal database for everything.** Commun. ACM, 49(1):88–95, 2006.

[Gifford91] Gifford, D., Jouvelot, P., Sheldon, M. and O'Toole, J. **Semantic File Systems**. *13th ACM Symposium on Principles of Programming Languages*, October 1991.

[Gonçalves04] Gonçalves, D. and Jorge, J. **Telling Stories to Computers**. In *Proceedings CHI2004*, ACM Press, 27-29 April 2004, Vienna, Austria.

[Malone83] Malone, T. **How do People Organize their Desks? Implications for the Design of Office Information Systems**, *ACM Transactions on Office Information Systems*, 1(1), pp 99-112, ACM Press 1983.

[Porter80] Porter, M. F. **An algorithm for suffix stripping**. *Program* 14, 130-137, 1980.

[Rekimoto99] Rekimoto, J. **Time-machine computing: a time-centric approach for the information environment**. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, pages 45-54, ACM Press, 1999.

[Salton88] Salton, G. **Automatic text processing**, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1988.

[Shamir04] Shamir, A.. **A View on Views**. In *Proceedings SmartGraphics04*, Banff Center, Canada, May 2004.

[Whittaker96] Whittaker, S., Sidner, C. **Email overload exploring personal information management of email**. In *Conference proceedings on Human factors in computing systems*, pages 276-283, ACM Press, 1996.