Terceira Geração de Sistemas de Pesquisa de Informação

João Ferreira
Instituto Superior de Engenharia de Lisboa
jferreira@deetc.isel.ipl.pt

Rui Jesus
Instituto Superior de Engenharia de Lisboa
rmfj@isel.ipl.pt

Arnaldo Abrantes
Instituto Superior de Engenharia de Lisboa
aja@cedet.isel.ipl.pt

Sumário: Pretende-se discutir e fundamentar um conjunto de ideias necessárias para o desenvolvimento de sistemas de pesquisa com o objectivo de melhorar o desempenho dos mesmos. Assuntos como a personalização, perfil, interfaces de ajuda, uso de sistemas de classificação e contextualização da pesquisa são analisados e discutidos numa abordagem que permite integrar de uma forma unificada as suas potencialidades.

Palavras-chave: Pesquisa de Informação, Sistema de Pesquisa, Personalização.

1 Introdução

Desde o aparecimento da Internet que o problema do excesso de informação e da respectiva recuperação tem sido abordado. Tornou-se prática comum a construção de sistemas de pesquisa (e.g. Altavista, Yahoo, Google). O objectivo destes sistemas é, dada uma necessidade de informação de um utilizador expressa numa pergunta por um conjunto de termos que o utilizador considere descreverem as suas necessidades, que devolva um conjunto de documentos. Estes sistemas podem dividir-se em duas grandes classes: os que trabalham num espaço aberto (e.g. Internet) e os que trabalham num espaço fechado, com colecções específicas e perguntas previamente elaboradas para as quais se conhece o conjunto de documentos relevantes. Historicamente podem dividir-se os sistemas em duas gerações:

- Os iniciais, atingindo o seu expoente máximo nos motores comerciais com o Altavista, o Excite e o Lycos (1992-1997), retirando-se apenas informação textual dos documentos, sendo posteriormente comparada;
- Segunda geração, que começou com a abordagem introduzida pelo Google (desde 1998), com base no seguimento das ligações dos documentos.

Pretende-se discutir os requisitos essenciais para a próxima geração, aqui denominada como 3ª geração de sistemas de pesquisa, nos quais a personalização assume um papel fundamental na opinião dos autores.

Personalização

Personalização significa a existência de uma base de dados para guardar o perfil do utilizador e um conjunto de definições locais. Dada a falta de um sistema global para tratar este assunto e para evitar os problemas que um tal sistema originaria (e.g. privacidade, segurança), a melhor abordagem para este problema é solucioná-lo do lado do cliente, através de uma nova geração de *browsers*, capaz de guardar e manipular a informação dos utilizadores.

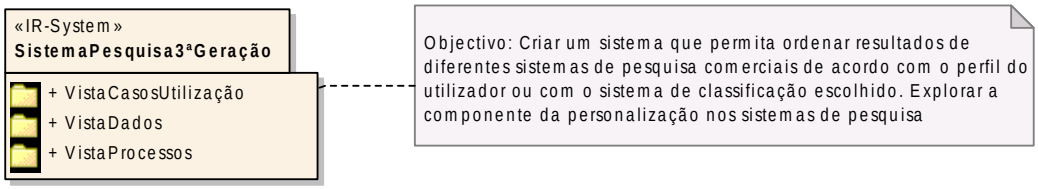

Figura 1: Sistema de pesquisa de 3ªgeração.

2 Vista de Casos de utilização

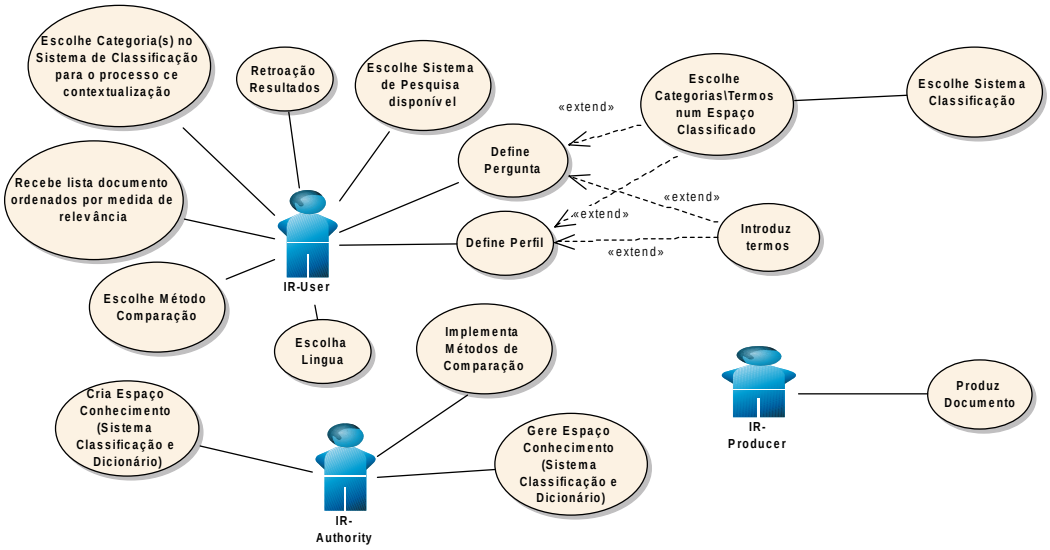

Figura 2: Vista de Caso de Utilização do sistema de pesquisa de 3ª geração.

Os ***IR-Actores*** são:

– ***IR-User***, para além do papel habitual que desempenha nos sistemas de pesquisa pode ainda escolher os sistemas de pesquisa comerciais disponíveis, definir o perfil, escolher os métodos de comparação, escolher a(s) categoria(s) do sistema de classificação para o processo de contextualização, de forma a definir no sistema qual o contexto que pretende e a linguagem com que quer fazer a pesquisa de informação;

– ***IR-Authority***, gere o espaço de conhecimento (sistemas de classificação e dicionários). Implementa métodos de comparação;

– ***IR-Producer***, produz documentos, que se encontram disponibilizados na *Web*.

3 Vista de dados

Pergunta, conjunto de termos representativos das necessidades de informação do utilizador que pode ser expressa pela introdução livre de termos ou pela navegação num espaço classificado apropriado. O utilizador escolhe a língua em que a pergunta se encontra e a língua em que deseja transformar a pergunta. Escolhe também o conjunto de sistemas de pesquisa dos quais quer obter resultados. A pergunta necessita de uma interface de ajuda ao utilizador que consiste num conjunto de ferramentas para ajudar o utilizador a formular

correctamente as perguntas, corrigindo erros ortográficos, indicando sinónimos de termos (através do uso de dicionários), permitindo a navegação em sistemas de classificação temáticos, relembrado perguntas feitas no passado. Esta interface permite também a pesquisa em diferentes línguas, através do uso de um sistema central de tradução

Dicionário, auxilia a elaboração da pergunta livre, evitando eventuais erros ortográficos. O dicionário foi implementado em duas versões: uma para a língua inglesa, de *Roger Mitton da Oxford Advanced Learner* http://www.oup.com/elt/global/products/oald/ e outro para a língua Portuguesa uma versão simplificada do dicionário electrónico da Porto Editora.

ResultadoSP, é a lista ordenada de documentos, por ordem de relevância, que o sistema de pesquisa (SP) escolhido considerou.

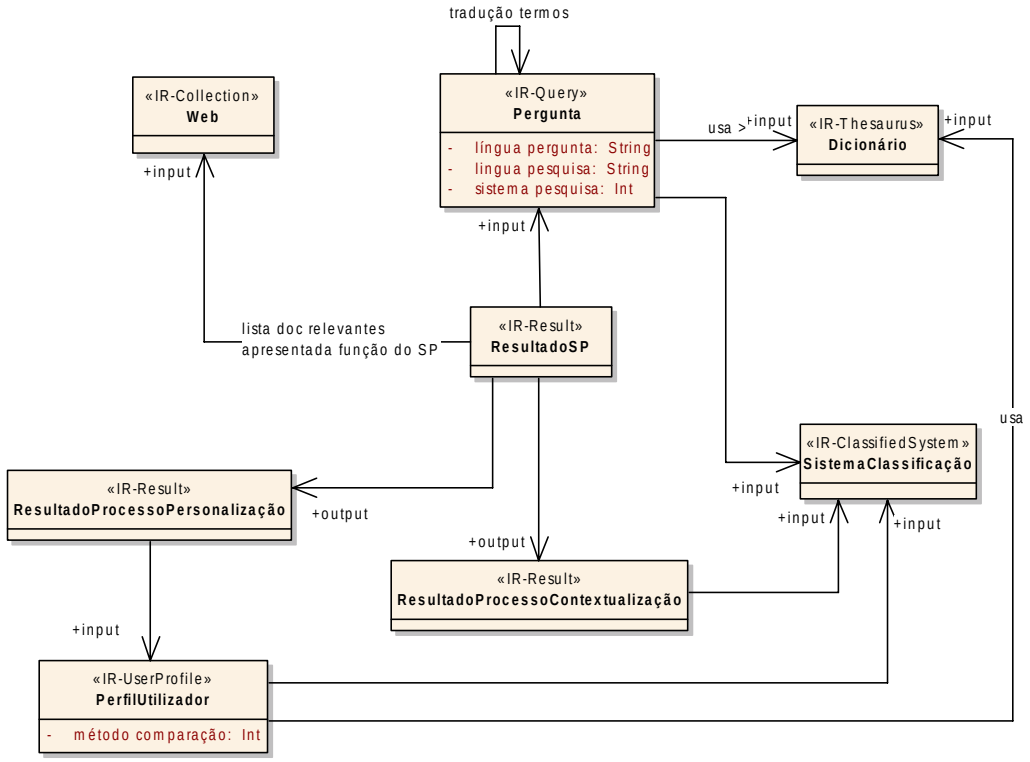

Figura 3: Vista de dados do sistema de pesquisa de 3ª geração

Sistema de Classificação, são disponibilizados sistemas gerais (Yahoo) e específicos (ACM, MSC). O sistema de classificação no processo de contextualização serve de *input* para que os resultados sejam ordenados de acordo com o espaço de conhecimento representado no sistema de classificação. O sistema de classificação é usado para sugerir ao utilizador termos para as perguntas bem como para o perfil e simultaneamente o agrupamento de termos dos documentos considerados relevantes permite identificar as classes às quais corresponderiam os temas no espaço classificado. Esta correspondência nem sempre tem sucesso, havendo necessidade da intervenção humana quando não existe semelhança entre os termos dos documentos e os do espaço classificado.

ResultadoProcessoPersonalização, são os resultados obtidos pelo processo de comparação escolhido, o qual compara os índices dos documentos considerados relevantes pelos diferentes SP escolhidos, com o perfil do utilizador.

PerfilUtilizador, contem termos representativos dos interesses estáveis obtidos por introdução de termos (assistida por corrector ortográfico) ou então por navegação no espaço classificado. A criação de um perfil, para evitar as questões de privacidade, requer uma base de dados local capaz de guardar a informação referente aos utilizadores (e.g., perfil, conteúdos de personalização). Estes dados são guardados localmente no cliente num formato que possa ser interpretado pelo sistema de pesquisa (lado do servidor) ou localmente pelo processo de personalização;

Os dados locais no cliente são: (1) pergunta, (2) perfil, (3) resultadoprocessopersonalização, (4) resultadoprocessocontextualização. Os dados centrais são: (1) sistemas de classificação; (2) dicionários disponíveis. Na *Web*, temos a colecção de documentos e os resultados do sistema de pesquisa.

4 Vista de Processos

Os principais processos são:

- **Tradutor**, recebe os termos, que vai traduzir tendo com referências a língua em que foi formulada a pergunta e da língua pretendida;
- **Processo de pesquisa**, corresponde ao processo padrão, podendo variar os processos de indexação, comparação e optimização consoante o sistema escolhido;
- **Processos de indexação**, os documentos identificados como relevantes são indexados pelo processo padrão de indexação;
- **Processo de personalização**, usa informação local para reordenar a informação a apresentar ao utilizador, pelo uso do perfil. Este processo usa o índice dos documentos considerados relevantes e por meio de comparação (escolhido pelo utilizador) os termos dos representativos dos documentos são comparados com os do perfil do utilizador. Deste processo resulta um menor número de documentos identificados como relevantes. Este processo pode também disponibilizar informação do perfil existente e usa processos de retroacção automáticos e manuais para ajustar os termos e as medidas existentes no perfil;
- **Contexto de pesquisa**, usa os termos da(s) categoria(s) do sistema de classificação escolhido para reduzir o número de documentos identificados como relevantes pelos diferentes sistema de pesquisa. O processo compara o índice dos documentos identificados como relevantes com os termos da(s) categoria(s).

Os processos estabelecidos do lado do cliente, originam a necessidade de um sistema central, capaz de gerir e implementar um conjunto de sistemas de classificação inerente às diferentes áreas do conhecimento, disponibilizar de forma uniforme dicionários em diferentes línguas bem como permitir a utilização de um

sistema central de tradução de termos associado a diferentes contextos.

O sistema local tem disponíveis os seguintes métodos de comparação (Disponibilizados centralmente): método vectorial (lnu-ltc), probabilísticos (fórmulas BMxx), seguimento de ligações, modelos linguísticos e também combinações.

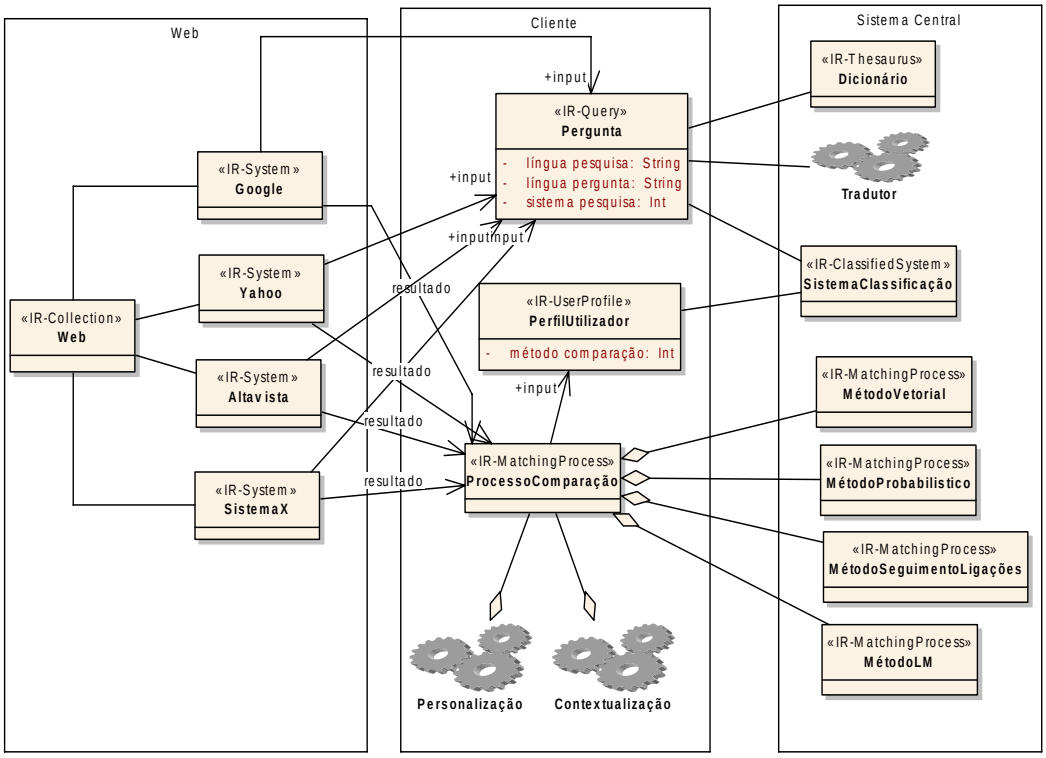

Figura 4: Vista de processos do sistema de pesquisa de 3ª geração proposto.

5 Conclusões

O sistema descrito encontra-se em fase de construção, e brevemente obter-se-ão resultados. O presente trabalho pretende mostrar uma reflexão sobre as direcções a tomar no que se refere a sistemas de recuperação de informação. Torna-se fundamental explorar as potencialidades dos perfis, sendo interessante a manipulação do lado do cliente para evitar a problemática associada aos temas privacidade e segurança e à dimensão de uma base de dados com os perfis dos utilizadores.

Referências

[1] Ferreira J, Silva A, Delgado J. (2005). Modelação da Pesquisa de Informação. JET2005.

[2] http://www.oup.com/elt/global/products/oald/