Linguagem para Modelação de Sistemas de Pesquisa de Informação

João Ferreira
ISEL
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
Jose.Delgado@tagus.ist.utl.pt

Sumário: O presente trabalho pretende abordar o problema da falta de uniformização de conceitos, fórmulas e parâmetros na área da pesquisa de informação, introduzindo uma linguagem própria com base nos mecanismos de extensão do UML, a qual serve de base à construção de modelos abstractos para a PI. Este modelos constituem um conjunto de bibliotecas cuja integração numa infra-estrutura permite construir sistemas de pesquisa de informação de uma forma simplificada, modular e uniforme..

Palavras chave: Pesquisa Informação, Modelação, Linguagem, UML.

1 Introdução

A pesquisa de informação tem-se desenvolvido explorando as propriedades estatísticas dos documentos, introduzindo um conjunto de simplificações ad-hoc, as quais são validadas em ambientes de teste. Não existem padrões nem uma teoria geral que possa contextualizar o problema. O trabalho desenvolvido neste domínio tem sido orientado para a criação de métodos e sistema de pesquisa, não existindo uniformização de notações, nem uma linguagem própria para a descrição dos problemas associados. Neste trabalho pretende-se dar os primeiros passos nesta área, usando as potencialidades do UML. Este trabalho insere-se no objectivo da construção automática de sistema de PI a partir de um conjunto de necessidades específicas de grupos de utilizadores. A perspectiva geral na qual o trabalho se integra encontra-se definida na Figura 1, da qual iremos abordar neste artigo a definição da IR-Language, língua criada com base nos mecanismos de extensão do UML, adaptada às necessidade especificas da PI.

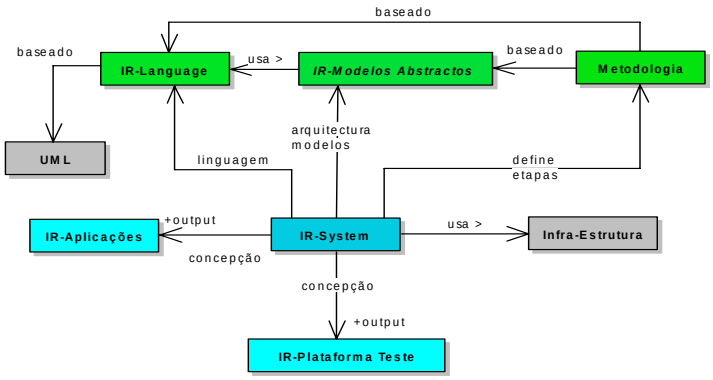

Figura 1: Etapas para a construção automatica de sistemas de PI.

Para facilitar a modelação do problema da construção de sistemas de Pesquisa de informação são propostas vistas, as quais têm como objectivo facilitar o processo da construção de um sistema, permitindo uma visão parcial do todo. Têm um papel semelhante às diferentes vistas de um plano de construção de uma casa. São propostas três vistas, de acordo com a Figura 2: (1) IR-UseCaseView define os actores (IR-Actors),

apresentando uma sequência de acções que estes realizam no sistema de forma a obterem um resultado particular. Esta vista define as relações do sistema com o exterior bem como define os objectivos do sistema; (2) IR-DataView, define os dados de entrada e saída do sistema, sendo estes caracterizados por um diagrama de classe e uma sequência de acções. Nesta vista pretende-se caracterizar os dados e o seu respectivo fluxo; (3) IR-ProcessView, define uma sequência, os atributos e as operações necessárias a um conjunto de processos, para transformar os dados de entrada no resultado a apresentar ao utilizador.

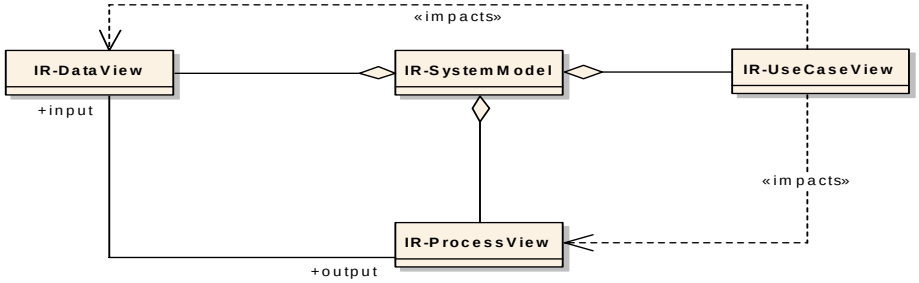

Figura 2: Vistas de representação de sistemas da linguagem para PI.

2 Vista de casos de utilização

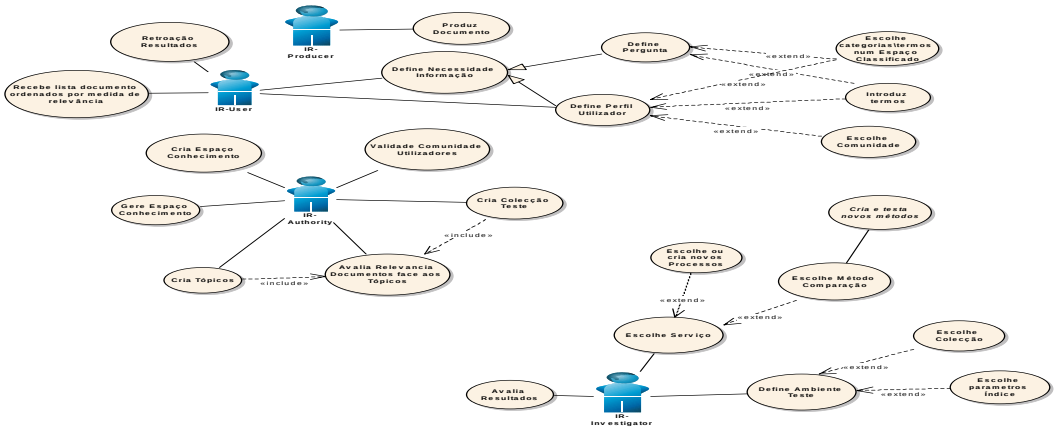

Figura 3: Vista dos casos de uso de um sistema de Pesquisa de Informação.

IR-Actor, caracteriza o actor do sistema o qual pode representar quatro papéis principais; (1) **IR-Autor (*IR-Producer*)**, o produtor de informação, usando os meios disponíveis para publicar a sua informação; (2) **IR-Utilizador** (***IR-User***) aquele que tem necessidade de recuperar informação e para o efeito expressa a sua necessidade por um conjunto de termos e espera que o sistema devolva uma lista ordenada de documentos relevantes; A necessidade de informação pode ser livre; (3) **IR-Autoridade** (***IR-Authority***) é responsável pela criação e gestão do espaço de conhecimento e simultaneamente pode identificar (criar) colecções de teste; (4)**IR-Investigador *(IR-Investigator)***, usa um sistema para testar algoritmos e abordagens de forma a contribuir para o avanço da ciência relacionada com a Pesquisa de informação. É responsável pela avaliação dos resultados obtidos.

3 Vista de Dados

***IR-Document* (Documento),** é a informação produzida pelo autor, não-estruturada, existente nos mais diversos formatos e tendo inerentes os problemas da subjectividade e do contexto da linguagem Humana.

***IR-Collection* (Colecção)** representa a fonte de informação para o sistema, constituída por um conjunto de documentos arquivados. A maior colecção existente é a *Web*. Existem diversas colecções construídas à medida para testes de sistemas. Numa colecção existe uma grande variedade de formatos, tamanhos de documentos, temas/assuntos. Uma colecção pode dividir-se em várias sub-colecções. As colecções podem ser armazenadas de uma forma centralizada ou distribuída.

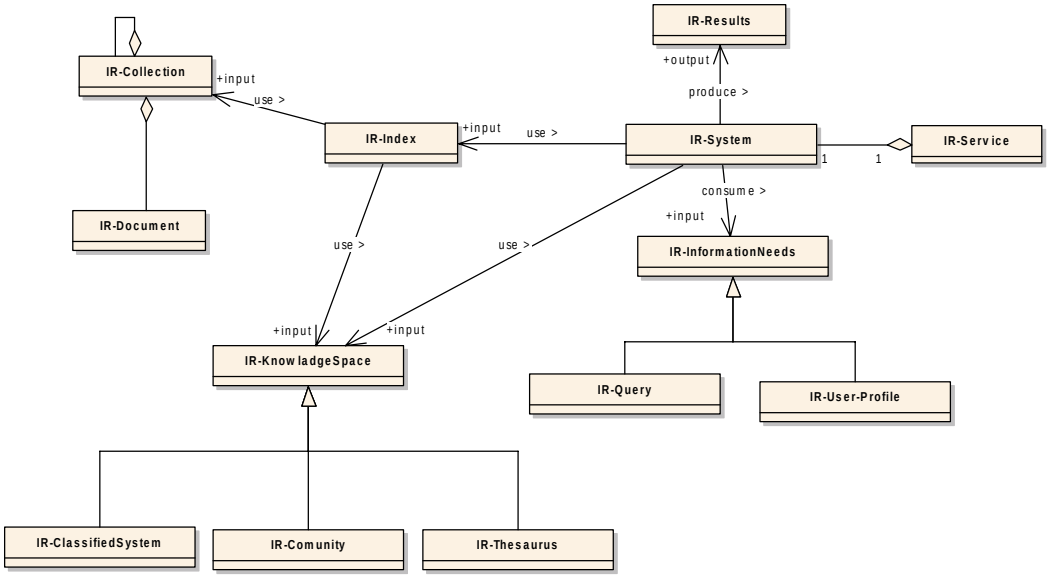

Figura 4: Vista dos perfis relacionados com a vista dos dados

***IR-Index* (Índice),** é o resultado da operação de criação de um representativo, de menores dimensões, de uma colecção. Os representativos encontram-se arquivados numa base de dados apropriada. É constituído essencialmente por termos representativos dos documentos com as respectivas frequências e baseado nas propriedades estatísticas dos documentos. É proposto um índice mais geral (não tão rápido) o qual pode servir para todos os métodos de Pesquisa. O Índice constitui a 'matéria-prima' para o funcionamento de um sistema de Pesquisa sendo previamente construído.

***IR- UserInformationNeeds* (Necessidade de Informação do Utilizador),** representa os interesses específicos de informação de um determinado utilizador, expresso por um conjunto de termos escolhidos pelo utilizador ou então pela navegação num espaço de conhecimento apropriado. Estas necessidades podem ser divididas em duas grandes classes: (1) ***IR-UserProfile* (Perfil Utilizador)**, representa os interesses estáveis de um utilizador. Pode ser formado por um conjunto de termos ou então por pontuação (identifica o atributo nota) dada a determinados eventos. Identifica a periodicidade com que o utilizador pretende receber a informação. O perfil contém ainda informação que identifica o utilizador do ponto de vista do sistema (endereço correio electrónico e login) e adicionalmente pode ter um Perfil negativo que reflecte temas nos quais o utilizador não está interessado em receber informação; (2) I**R-Query (Pergunta)**, representa o interesse momentâneo de um determinado utilizador, expresso através de um conjunto de termos. Estes termos são posteriormente trabalhados de forma a melhorar o desempenho de um determinado sistema.

IR-KnowladgeSpace (Espaço Conhecimento), representa o espaço organizado e previamente trabalhado por um conjunto de entidades. Este espaço é dividido em três grandes áreas: sistema de classificação, *thesaurus*/dicionários e comunidades de utilizadores (definições apresentadas em [2])

***IR-Results* (Resultados)**, é o resultado do serviço em causa, consistindo habitualmente numa lista de documentos ordenada por medida de relevância.

***IR-System* (Sistema),** é um conjunto integrado de recursos (humanos e tecnológicos) cujo objectivo é satisfazer adequadamente a totalidade das necessidades de um determinado serviço

***IR-Service* (Serviço)**, representa a generalização do conceito de sistema orientado para um determinado objectivo, do ponto de vista das acções a executar tendo em conta os objectivos definidos para os utilizadores. O sistema é constituído por um conjunto de acções, enquanto que o serviço está orientado para o conceito.

4 Vista de processos

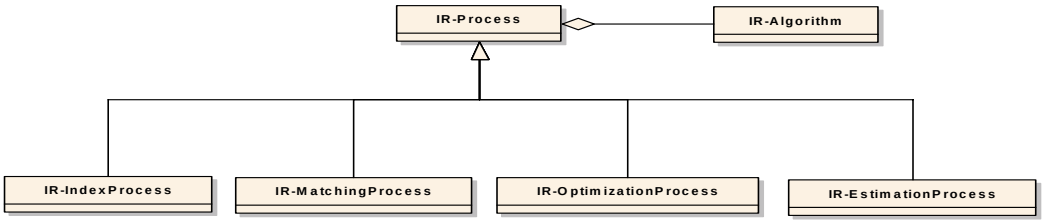

Figura 5: Vista dos processos principais de Pesquisa de informação.

***IR-Process* (Processo),** é um conceito vasto, que pretende designar uma sequência de actividades (agrupadas em fases e tarefas) executadas de forma sistemática e uniformizada, por intervenientes com responsabilidades bem definidas, e que a partir de um conjunto de entradas produzem um conjunto de saídas. Existem diversos processos, dos quais se realçam quatro específicos, ficando os restantes identificados como processos:

- ***IR-IndexProcess* (processo de indexação),** responsável por criar representativos dos documentos existentes numa colecção e é um dos principais processos do serviço de Pesquisa. O objectivo deste processo é criar um representativo do documento com dimensões inferiores. Os processos de indexação são orientados para o método de comparação a implementar. Para descrição mais detalhada ver [1];
- ***IR-MatchingProcess* (processo comparação),** por meio de um conjunto de algoritmos compara os representativos dos documentos com os representativos das necessidades de informação dos utilizadores resultando numa lista de documentos ordenados por ordem de relevância ou de acordo com uma medida previamente estabelecida. Cada um destes métodos encontram-se descritos em [2];
- ***IR-OptimationProcess* (processo optimização),** tem como objectivo melhorar a lista de documentos considerados relevantes, a apresentar aos utilizadores. Estão divididos em dois tipos principais: (1) os de retroacção, que trabalham os *inputs* do sistema (necessidades de informação e índice); (2) os de combinação que trabalham os resultados obtido. Para maior detalhe ver [1];

– ***IR-EstimationProcess*** **(processo estimar),** que a partir de colecções de teste, estimam parâmetros para modelos linguísticos, usados nos algoritmos de classificação.

5 Perfil UML para IR

Nas secções anteriores foram definidos estereótipos, os quais definem o perfil UML para a Pesquisa de informação. A Tabela 1 estabelece as relações (C-Criação; V-Validação; U-Uso; I-Mehoramentos) entre os diferentes estereótipos identificados.

Relações C - Cria; V - Valida; U - Usa; A- Avaliação; I-Melhora	IR-Actor	IR-Autor	IR-User	IR-Authority	IR-Investigator	IR-Document	IR-Collection	IR-Process	IR-IndexProcess	IR-OptimizationPro	IR-EstimationProc	IR-MatchingProce	IR-Index	IR-InformationNeeds	IR-Query	IR-UserProfile	IR-KnowladgeSpace	IR-Dicionary	IR-ClassifiedSyster	IR-Comunity	IR-System	IR-Service	IR-Results
IR-Actor																							
IR-Autor						C																	
IR-User															C	C					U	U	U
IR-Authority							C								C			C	C	V			
IR-Investigator							U	C	C	C	C	C	C								U	U	A
IR-Document		C							U				U								U		
IR-Collection				C	U																U		
IR-Process					C																U		
IR-IndexProcess					C	U							C					U	U		U		
IR-OptimizationProcess					C								I		I	I					U		I
IR-EstimationProcess					C							I									U		
IR-MatchingProcess					C						I		U		U	U					U		C
IR-Index					C	U			C	I		U						I	I		U		
IR-InformationNeeds																							
IR-Query			C	C						I		U						I	I	I	U		
IR-UserProfile			C							I		U						I	I	I	U		
IR-KnowladgeSpace																							
IR-Dicionary				C					U				I		I	I					U		
IR-ClassifiedSystem				C					U				I		I	I					U		
IR-Comunity				V											I	I					U		
IR-System			U		U	U	U	U	U	U	U	U	U		U	U		U	U	U		C	C
IR-Service			U		U																C		
IR-Results			U		A					I		C									C		

Tabela 1: Relações entre os estereótipos definidos no perfil UML para a Pesquisa de informação.

6 Modelos abstractos

Com base na linguagem proposta, definem-se os modelos abstractos, os quais disponibilizam um conjunto de bibliotecas padrão, para o processo de criação de sistemas modelares de IR. Devido à elevada quantidade de informação a gerir e armazenar, procura-se que estes modelos assentem numa infra-estrutura, que disponibilize uma base de dados para guardar o índice dos documentos, Figura 1. A Figura 6, identifica os principais modelos abstractos, do ponto de vista dos dados. Os modelos abstractos do ponto de vista do processo estão abordados em [1]. A partir destes modelos outros podem ser derivados constitíndo um conjunto de bibliotecas disponiveis para a concepção e construção de sistemas de pesquisa de informação.

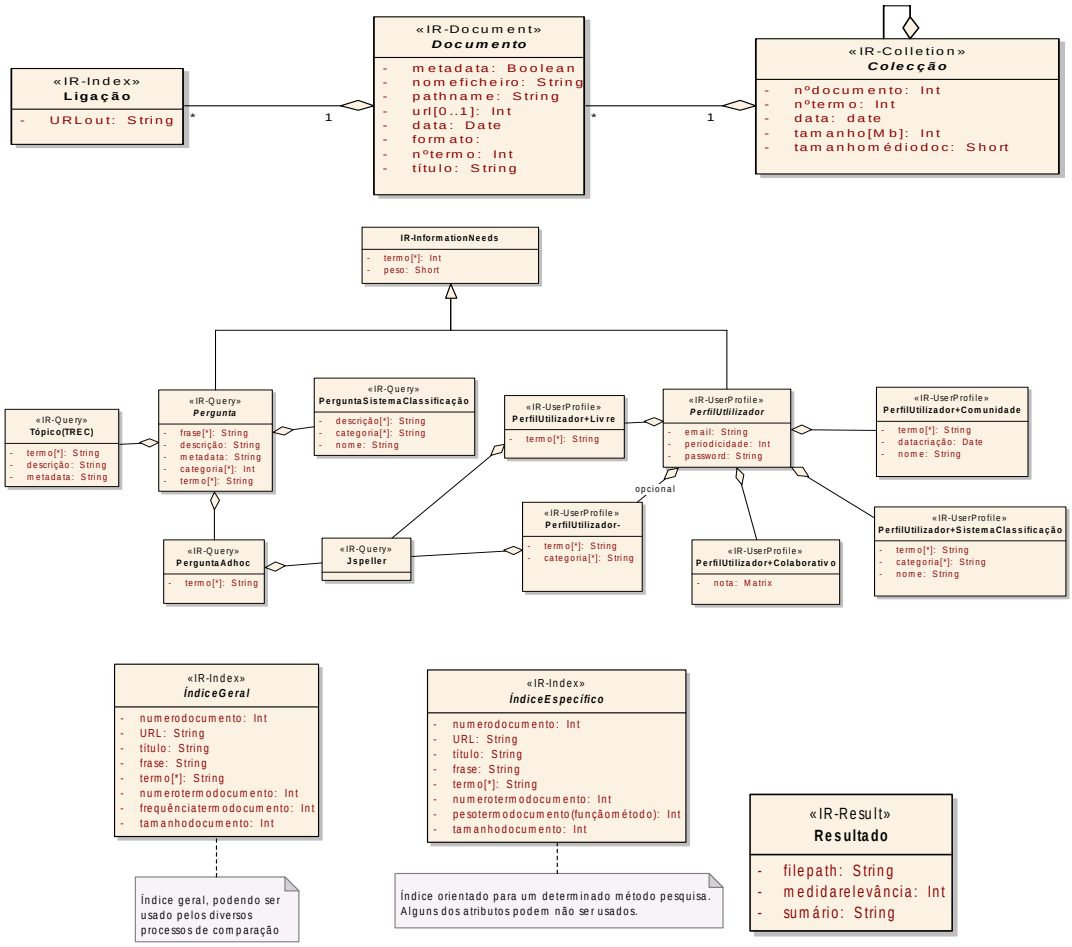

Figura 6: Modelos abstractos da vista de dados de um sistema de pesquisa.

Referências

[1] Ferreira J, Tese de Doutoramento. Metodologia para Concepção e Construção de Sistemas de Recuperação de Informação

[2] Ferreira J, Silva A, Delgado J. (2005). Métodos Estatísticos para Pesquisa de Informação. JET2005.

[3] Silva A, Videira C., (2005). UML - Metodologias e Ferramentas CASE (2ª Edição, revista e actualizada para o UML 2), ed. Centro Atlântico