andalucíaPeople

Un sistema de recomendación para sitios de ocio de Andalucía

Autor: Manuel Martín Salvador
Tutor: Juan Huete Guadix

Proyecto Fin de Carrera | Ingeniería Informática | Universidad de Granada

Introducción: motivación

2006

2007

2009

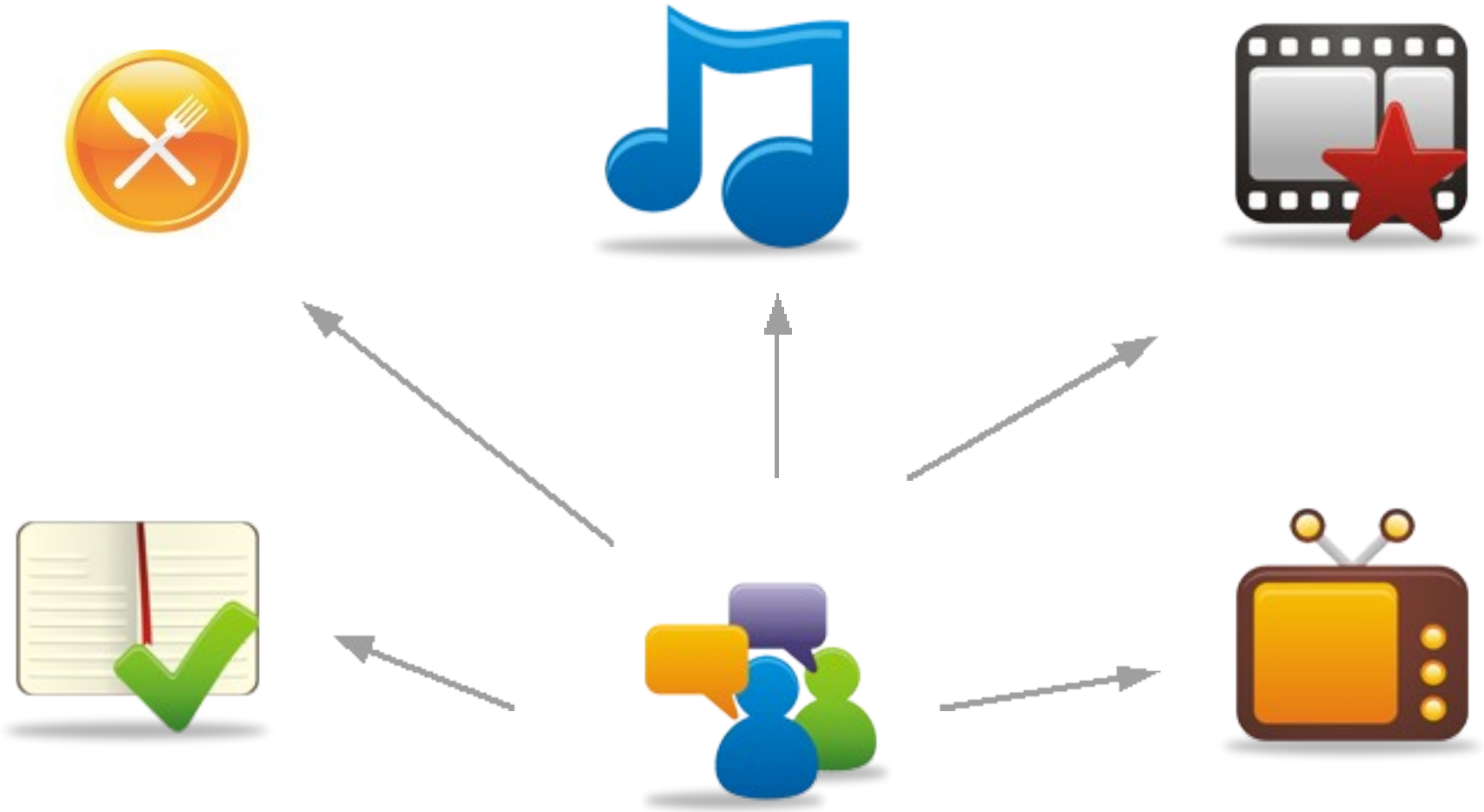
Introducción: situación inicial

- Base de datos con información de sitios
- Sistema de votación
- Sistema de comentarios y fotografías de los sitios
- Usuarios registrados
- Buscador
- Mapa
- Eventos y hoteles
- Amigos
- Mensajería privada

Introducción: mejoras planteadas

- Incluir las 8 provincias andaluzas
- **Sistema de recomendación**
- Incentivar al usuario
- Versión *mobile*
- Multi-idioma

Los sistemas de recomendación



Los sistemas de recomendación

Factores importantes

- Facilidad de uso

- Calidad de las recomendaciones

- Transparencia del sistema

Casos de uso

- Predicción

- Recomendación

Los sistemas de recomendación

Clasificación

- Sistemas de filtrado colaborativo

- Sistemas de recomendación basados en contenido

- Sistemas de recomendación híbridos

Técnicas de recomendación

- **Modelos probabilísticos**
- Árboles de decisión
- **Vecinos**
- Clustering
- Redes neuronales

Sistema de recomendación basado en contenido

Elementos involucrados

- Sitios (ítems) I_j — *Ej. Restaurante Wok*
- Etiquetas (descriptores) F_k — *Ej. Buffet, Japonés, Terraza...*
- Jerarquías de etiquetas C_u — *Música, Estilo, Instalaciones, Otros*
- Votos del usuario
- Conjunto de valoraciones — *R={0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5}*
- Perfil de usuario — Ej. Pop, Vegetariano, Wifi...
- Sitio a predecir para el usuario activo

Sistema de recomendación basado en contenido

Modelo probabilístico: red Bayesiana

- Grafo dirigido acíclico: nodos (variables) + arcos (relaciones)

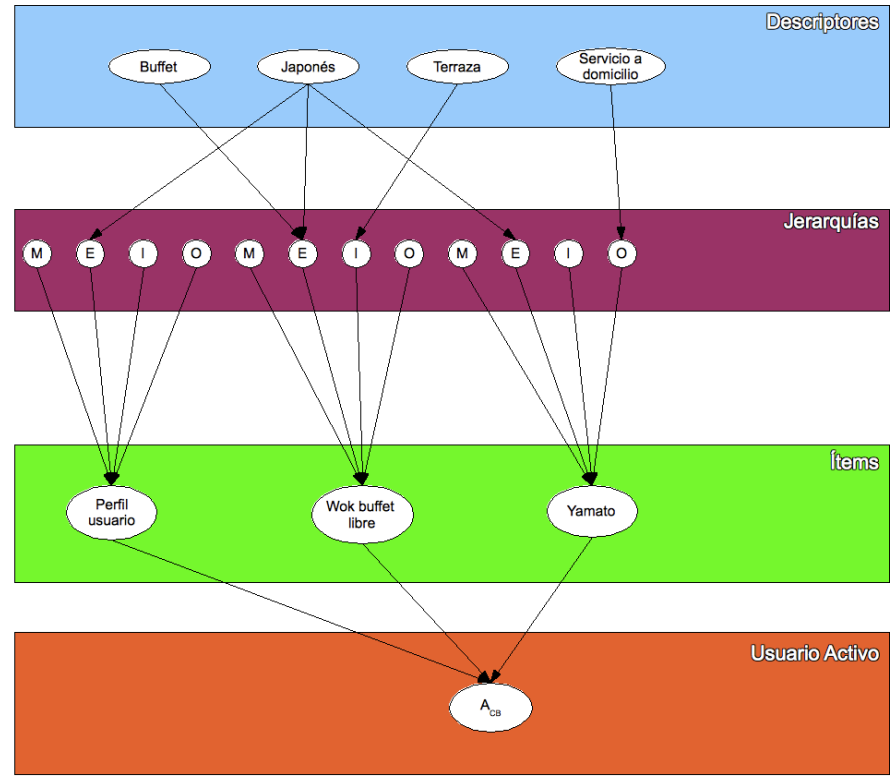
- En cada nodo $X_i \rightarrow$ distrib. de probabilidad condicional $Pr(X_i | Padres(X_i))$

Sistema de recomendación basado en contenido

Modelo probabilístico: red Bayesiana

- Grafo dirigido acíclico: nodos (variables) + arcos (relaciones)

- En cada nodo $X_i \rightarrow$ distrib. de probabilidad condicional $Pr(X_i | Padres(X_i))$



Sistema de recomendación basado en contenido

Modelo probabilístico: red Bayesiana

- Grafo dirigido acíclico: nodos (variables) + arcos (relaciones)

- En cada nodo $X_i \rightarrow$ distrib. de probabilidad condicional $Pr(X_i|Padres(X_i))$

Problema

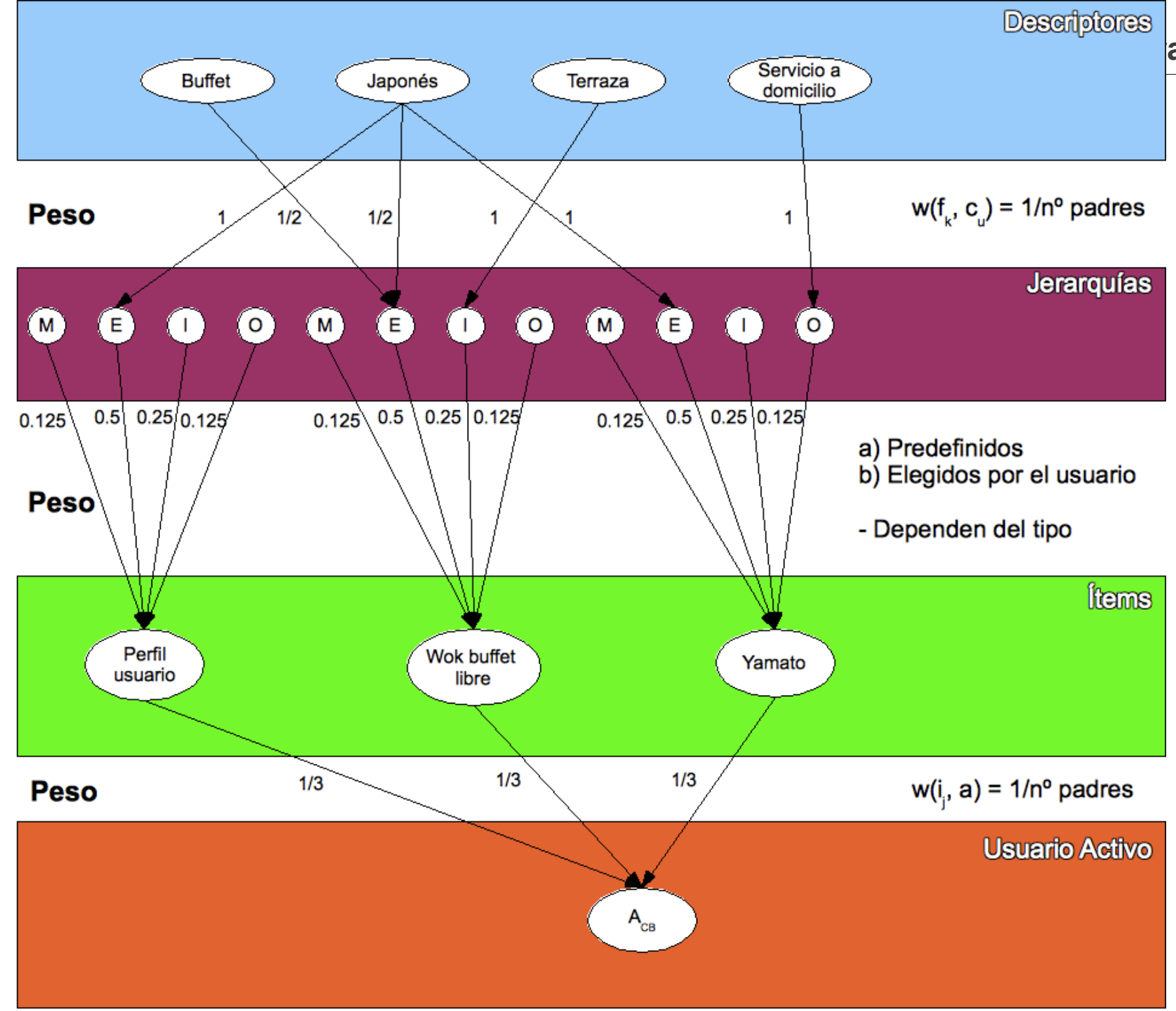
Ya que un nodo puede ser o no relevante, se necesitarían calcular 2^n distribuciones de probabilidad.
(*n* = número de padres)

Solución

Usar un modelo canónico de suma de pesos

$$Pr(x_{i,j}|Padres(X_i)) = \sum_{Y_k \in Padres(X_i)} w(y_{k,l}, x_{i,j})$$

Si



M=Música, E=Estilo, I=Instalaciones, O=Otros

Sistema de recomendación basado en contenido

Predicción del voto

- Problema: dado un sitio nuevo calcular su estimación del voto

Procedimiento

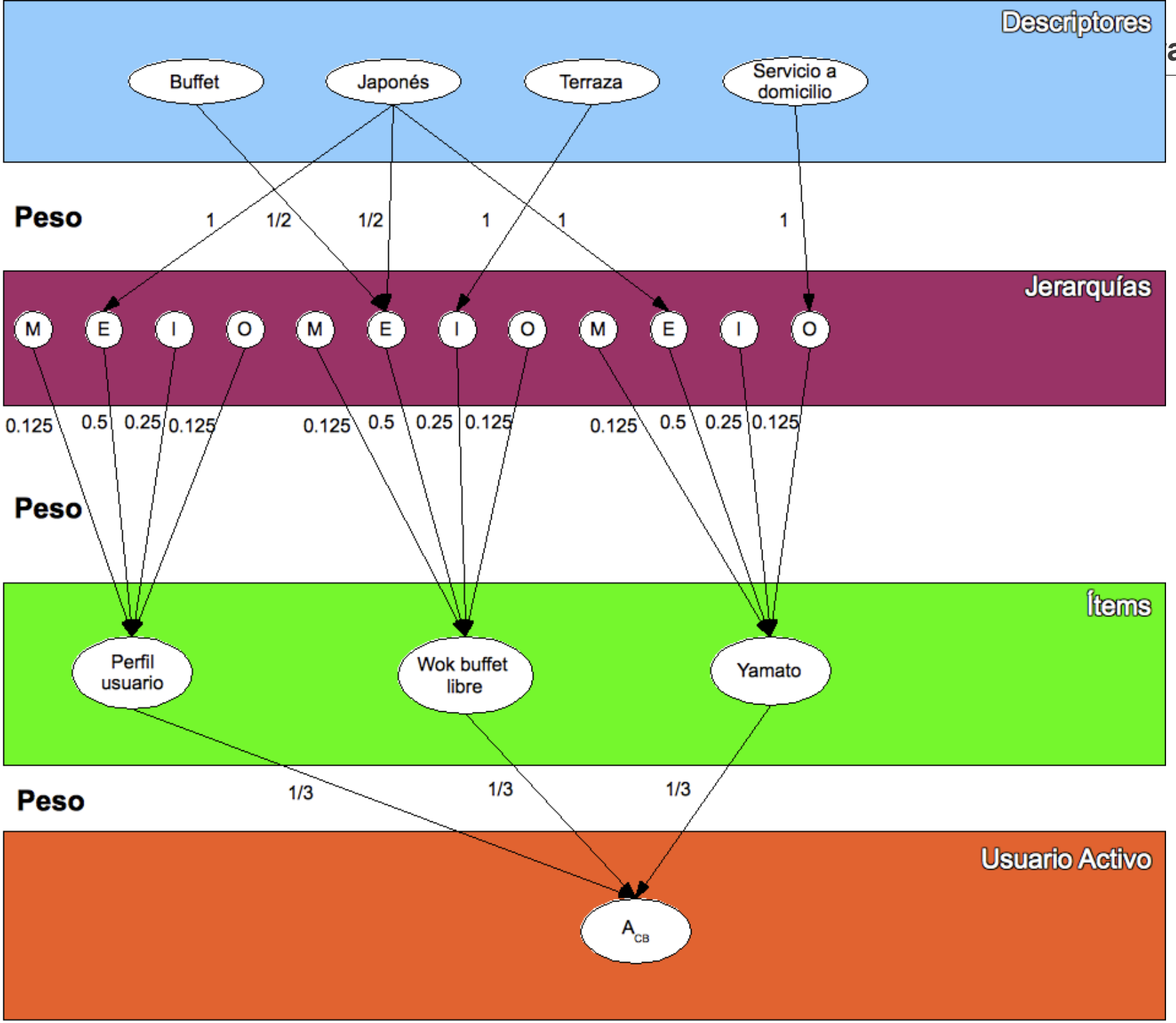
1º Se incluye el sitio en el modelo

2º Se propagan las probabilidades del sitio a sus etiquetas

3º Se hace una propagación top-down

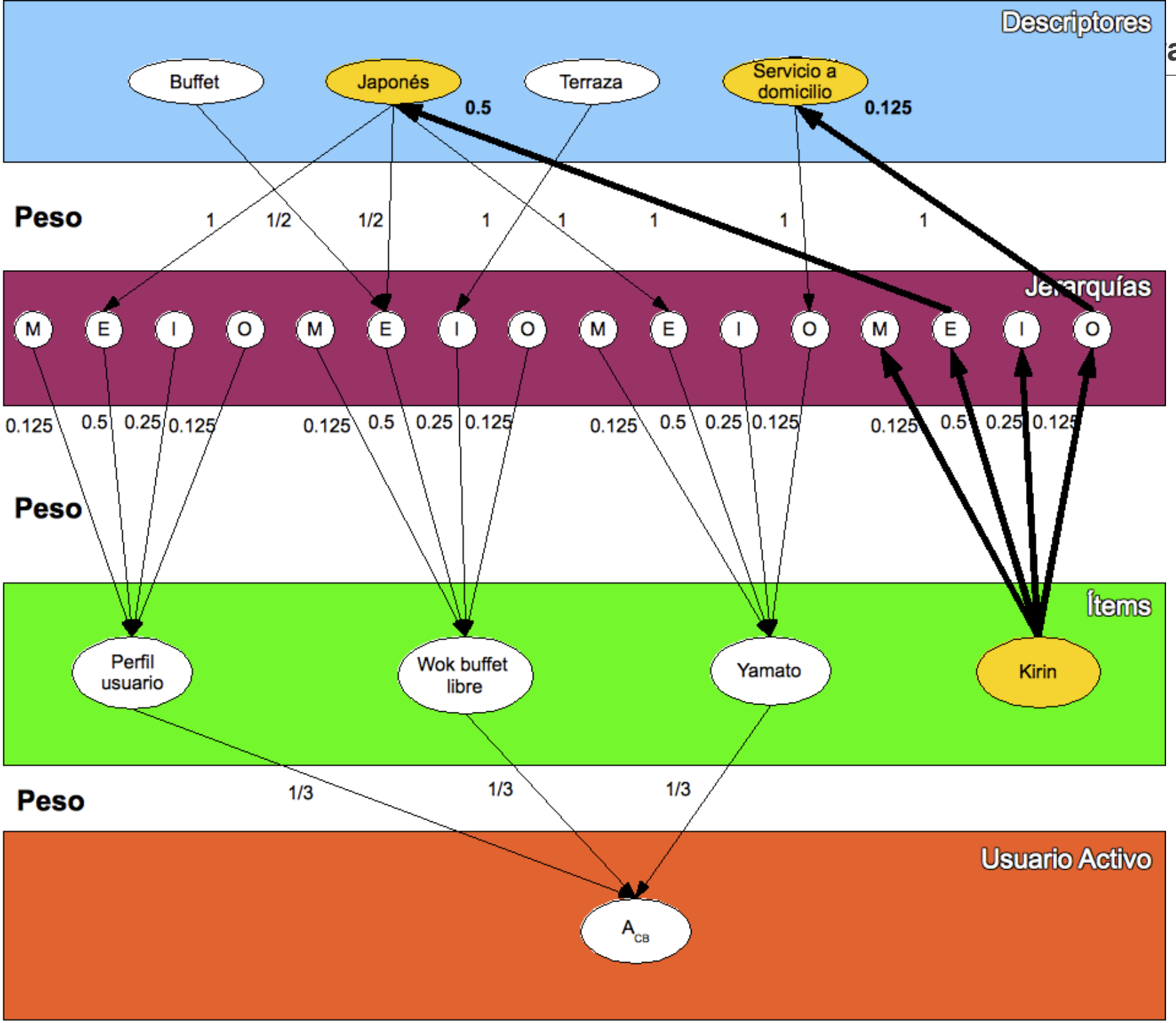
Se puede ver con más detalle en la memoria

Si



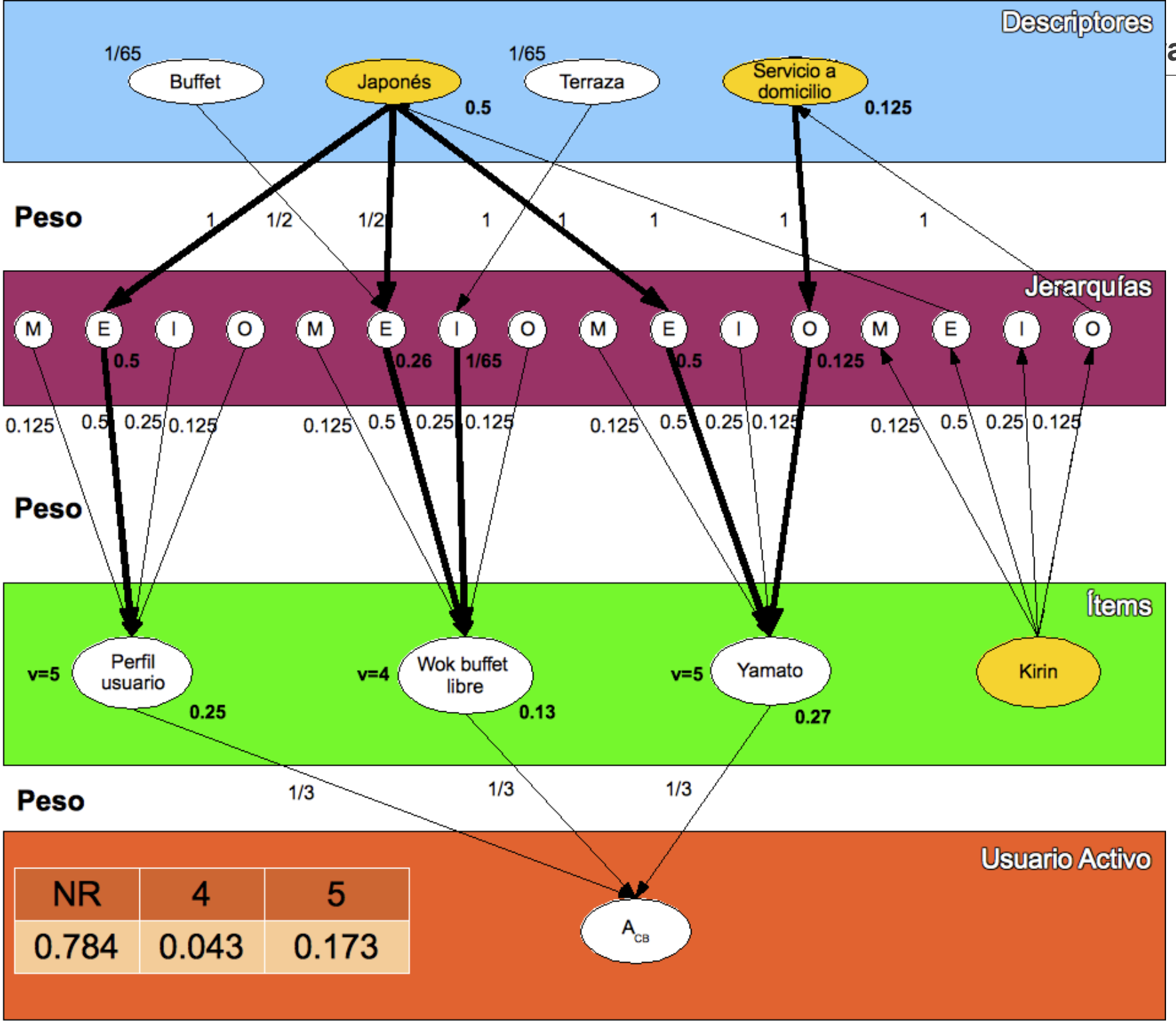
M=Música, E=Estilo, I=Instalaciones, O=Otros

Si



M=Música, E=Estilo, I=Instalaciones, O=Otros

Si



NR	4	5
0.784	0.043	0.173

M=Música, E=Estilo, I=Instalaciones, O=Otros

Sistema de recomendación basado en contenido

Predicción del voto

Valoración	NR	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Probabilidad	0.784	0	0	0	0	0	0	0	0.043	0	0.173
Normalizada	21.6%	0	0	0	0	0	0	0	0.199	0	0.801

- Voto promedio

$$predicción = \sum_{\forall s \in R} s \cdot Pr(A = s|ev)$$

Ej: 4.8

- Voto mediano

$$predicción = \{s | Pr(A < s|ev) \leq 0.5, Pr(A > s|ev) \geq 0.5\}$$

Ej: 5

- Voto máximo

$$predicción = \{s | Pr(A = s|ev) > Pr(A \neq s|ev)\}$$

Ej: 5

Sistema de recomendación basado en contenido

Limitaciones

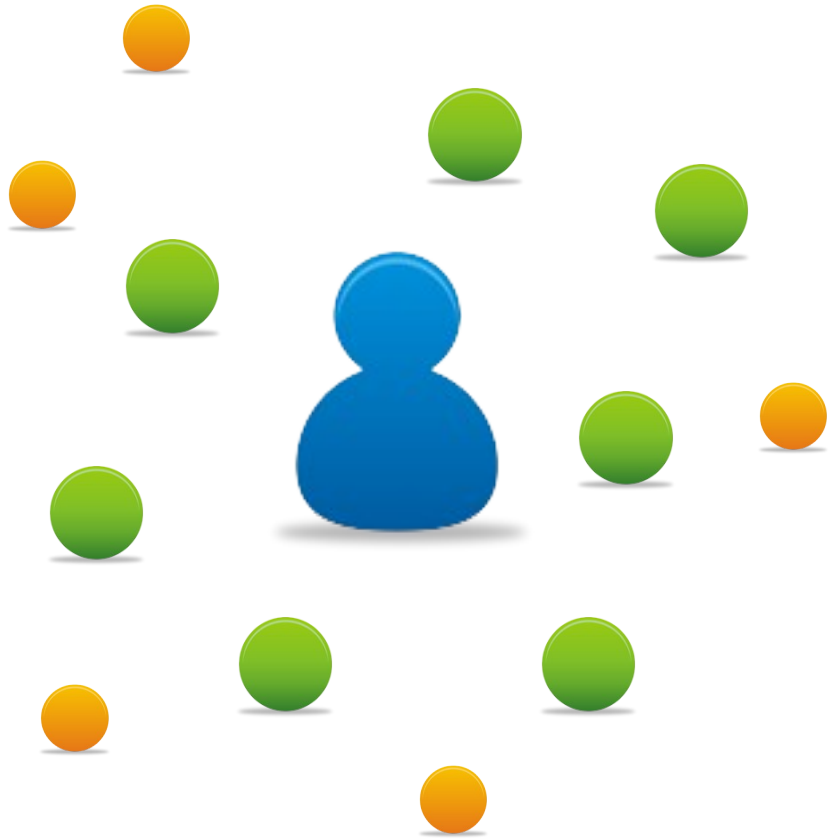
- Problema del *cold-starting*

  Estrategias de resolución:
  1. *Ignorar al usuario*
  2. *Tratarlo como el usuario medio*
  3. *Sitios populares*

- Recomendación de sitios mal valorados

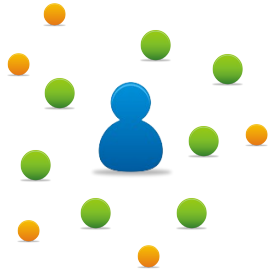
  Solución adoptada: *umbral*

Sistema de filtrado colaborativo



Vecinos más cercanos: aquellos usuarios más similares

Sistema de filtrado colaborativo



Procedimiento

Dado un sitio a predecir:

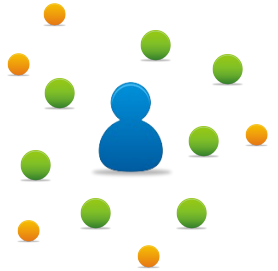
1º Calcular la similaridad del usuario activo con todos los usuarios que han votado el sitio

2º Seleccionar los *k* mejores

3º Predecir el voto, utilizando las valoraciones de los usuarios

$$predicción = \bar{r}_{u_a} + \frac{\sum_{h=1}^{k} sim(u_a, u_h) \cdot (r_{u_h, i_a} - \bar{r}_{u_h})}{\sum_{h=1}^{k} |sim(u_a, u_h)|}$$

Sistema de filtrado colaborativo



Medidas de similaridad

- Coseno
  COS

$$sim(u_x, u_y) = \frac{\sum_{h=1}^{m'} r_{u_x,i_h} \cdot r_{u_y,i_h}}{\sqrt{\sum_{h=1}^{m'} r_{u_x,i_h}^2} \cdot \sqrt{\sum_{h=1}^{m'} r_{u_y,i_h}^2}}$$

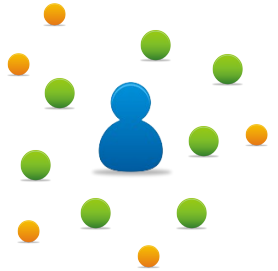
- Correlación de Pearson
  COR

$$sim(u_x, u_y) = \frac{\sum_{h=1}^{m'} (r_{u_x,i_h} - \bar{r}_{u_x}) \cdot (r_{u_y,i_h} - \bar{r}_{u_y})}{\sqrt{\sum_{h=1}^{m'} (r_{u_x,i_h} - \bar{r}_{u_x})^2} \cdot \sqrt{\sum_{h=1}^{m'} (r_{u_y,i_h} - \bar{r}_{u_y})^2}}$$

- Correlación de Pearson limitada
  CPC

$$sim(u_x, u_y) = \frac{\sum_{h=1}^{m'} (r_{u_x,i_h} - r_{med}) \cdot (r_{u_y,i_h} - r_{med})}{\sqrt{\sum_{h=1}^{m'} (r_{u_x,i_h} - r_{med})^2} \cdot \sqrt{\sum_{h=1}^{m'} (r_{u_y,i_h} - r_{med})^2}}$$

Sistema de filtrado colaborativo



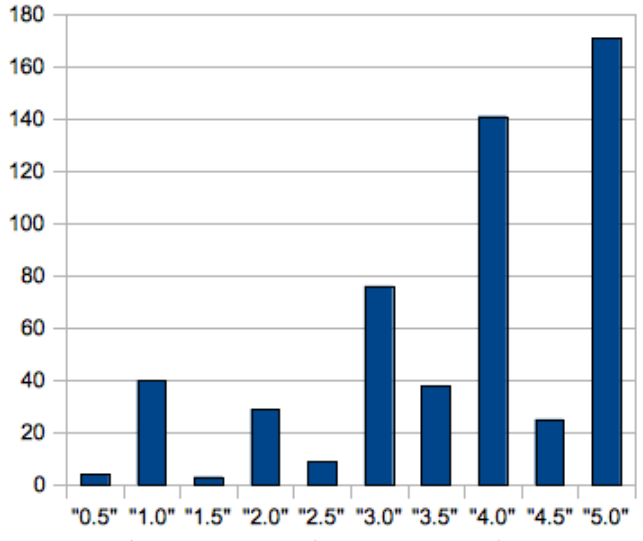
Limitaciones

- Problema del *cold-starting*:
  1. Usuario nuevo
  2. Sitio nuevo
  3. Comunidad nueva

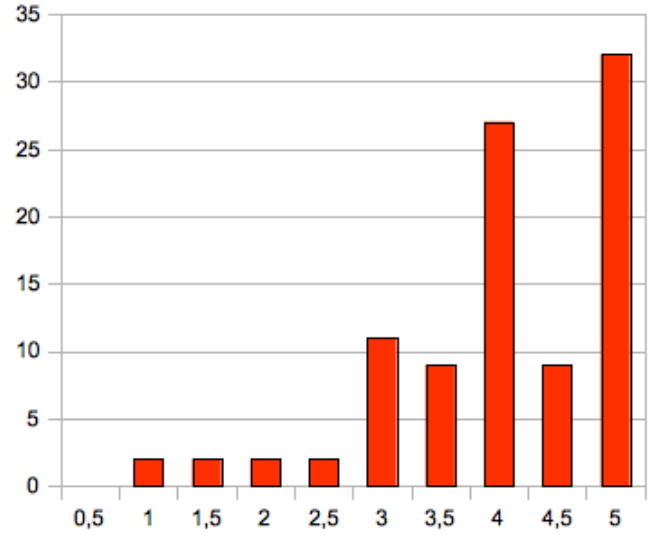
- Recomendaciones en otras ciudades

Evaluación: colección de datos

- **Usuarios registrados:** 330
- **Usuarios que han rellenado su perfil:** 27
- **Usuarios con más de 1 voto:** 54
- **Usuarios con más de 5 votos:** 27
- **Usuarios con más de 10 votos:** 15
- **Sitios:** 331
- **Votos:** 536



Votos



Votos medios

Evaluación: métricas y metodología

Métricas

- Error medio absoluto (MAE)

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

- Error cuadrático medio (MSE)

$$MSE = \frac{\sum_{i=1}^{N} |p_i - r_i|^2}{N}$$

- Porcentaje de predicción

Metodología

- Entrenamiento 80% + Test 20%

- *Leave one out*

Evaluación: resultados

Metodología leave one out

Método	MAE	MSE	%Predicción
Voto medio	0.87678	1.4085	92
Coseno (contenido)	1.49458	3.0523	11
Basado en contenido	0.9791	1.8858	27
Colaborativo COS	0.9735	1.7680	63
Colaborativo CPC	0.8938	1.3616	55
Colaborativo COR	0.9917	1.7019	36

Soluciones desarrolladas

1. Cambio de plataforma: *django*
2. Incentivos para el usuario
3. Sistema de recomendación
4. Jerarquización de etiquetas
5. Versión *mobile*
6. Soporte multi-idioma
7. Blog y *feeds*

Demo!

http://andaluciapeople.com