



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA  
INFORMÁTICA

DEPARTAMENTO DE ARQUITECTURA Y TECNOLOGÍAS DE  
COMPUTADORES

# HealthRecSys: Sistema Recomendador para la salud

TESIS DOCTORAL DE

CARLOS LUIS SÁNCHEZ BOCANEGRA

DIRECTORES

JOSE LUIS SEVILLANO RAMOS

LUIS FERNÁNDEZ LUQUE



# Dedicatoria

*Dedico este trabajo a mis padres que lo dieron todo por mí y  
-especialmente- a mi madre que me acompaña siempre.  
A mi familia que me apoya en todo momento -especialmente- a las dos  
mujeres más importante de mi vida que han estado a mi lado siempre.*



# Índice general

|   |          |
|---|----------|
| Índice de figuras   | 9        |
| Índice de tablas  | 11       |
| <b>1. Introduction</b>  | <b>1</b> |
| 1.1. Approach . . . . .   | 2        |
| 1.2. Thesis structure . . . . .   | 3        |
| 1.3. Objectives . . . . .   | 4        |
| 1.4. List of projects and publications . . . . .  | 5        |
| <b>2. Estado del arte</b>   | <b>7</b> |
| 2.1. Elementos partícipes en la recomendación . . . . .   | 8        |
| 2.1.1. Perfil de usuario . . . . .  | 8        |
| 2.1.2. Ítems . . . . .  | 8        |
| 2.1.3. Contexto . . . . .   | 9        |
| 2.1.4. Dominio . . . . .  | 10       |
| 2.2. Métodos básicos en los Sistemas Recomendadores . . . . .   | 11       |
| 2.2.1. Colaborativos (Collaborative methods) . . . . .  | 11       |
| 2.2.2. Recomendación del vecino más próximo basada en ítem (Item-based nearest neighbor recommendation) | 15       |
| 2.2.3. Métodos basados en contenidos (Content-based methods) . . . . .                                  | 16       |
| 2.2.4. Métodos basado en el conocimiento (Knowledge-based methods) . . . . .                            | 20       |
| 2.3. Otros métodos de recomendación más avanzados . . . . .   | 21       |

|           |   |           |
|-----------|---|-----------|
| 2.3.1.    | Matriz de factorización/Modelos de factor latente (Matrix factorization/latent factor models) . . . . . | 22        |
| 2.3.2.    | Minería de asociación de reglas (Association rule mining) . . . . .                                     | 22        |
| 2.3.3.    | Recomendaciones usando ontologías y semántica web   | 23        |
| 2.3.4.    | Métodos híbridos (Hybrid methods) . . . . .   | 24        |
| 2.4.      | Desafíos existentes en los sistemas recomendadores . . . . .  | 24        |
| 2.4.1.    | Evaluación implícita y explícita (ratings) . . . . .  | 24        |
| 2.4.2.    | Dispersión de los datos (data sparsity) . . . . .   | 25        |
| 2.4.3.    | Problema del arranque en frío (cold-start) . . . . .  | 25        |
| 2.4.4.    | Serendipia/Sobre-especialización (Serendipity/overspecialization) . . . . .                             | 26        |
| 2.4.5.    | Latencia . . . . .  | 26        |
| 2.5.      | Sistemas Recomendadores en Salud . . . . .  | 26        |
| 2.5.1.    | Sistemas recomendadores y el registro de salud personal (PHR) . . . . .                                 | 27        |
| 2.5.2.    | Sistemas recomendadores y Social Media . . . . .  | 29        |
| 2.5.3.    | Procesado de Lenguaje Natural . . . . .   | 30        |
| <b>3.</b> | <b>HealthRecSys: Recomendador en salud</b>  | <b>35</b> |
| 3.1.      | Contexto de la aplicación . . . . .   | 35        |
| 3.1.1.    | Saluteca y Diavideos . . . . .  | 36        |
| 3.2.      | Metodología de Evaluación . . . . .   | 48        |
| 3.2.1.    | Evaluación cualitativa de acuerdo - Kappa Cohen . .   | 49        |
| 3.2.2.    | Evaluación de las recomendaciones de MedlinePlus - <i>Precision@k</i> . . . . .                         | 49        |
| 3.2.3.    | Evaluación de las recomendaciones de MedlinePlus - <i>nDCG</i> . . . . .                                | 51        |
| 3.3.      | HealthRecSys . . . . .  | 52        |
| 3.3.1.    | Etapa previa - Etiquetas . . . . .  | 52        |
| 3.3.2.    | Etapa primera - Términos para recomendar . . . . .  | 54        |
| 3.3.3.    | Etapa segunda - Ontologías . . . . .  | 57        |
| 3.3.4.    | Etapa tercera - Procesado de Lenguaje Natural- NLP  | 61        |
| 3.4.      | Evaluación de HealthRecSys . . . . .  | 66        |

|  |            |
|--|------------|
| <b>4. Discussion</b>   | <b>73</b>  |
| 4.1. Related work . . . . .                                  | 73         |
| 4.1.1. Milestones achieved in the thesis . . . . .           | 75         |
| 4.2. Future works . . . . .                                  | 76         |
| <b>A. cTAKES configuration</b>                               | <b>79</b>  |
| <b>B. Introduction on Health Recommender Systems chapter</b> | <b>83</b>  |
| <b>C. HealthRecSys paper</b>                                 | <b>101</b> |
| <b>Bibliografía</b>  | <b>113</b> |





# Índice de figuras

|  |    |
|--|----|
| 2.1. Esquema de un perfil - Fuente berkeley.edu . . . . .  | 9  |
| 2.2. Ejemplo de un perfil de usuario - Fuente berkeley.edu . . .                                     | 10 |
| 2.3. Ejemplo de recomendaciones (ítems) para Diabetes . . . . .                                      | 11 |
| 2.4. Estudio del contexto en una recomendación - fuente Inter-<br>ation Designs Foundation . . . . . | 12 |
| 2.5. Sistema Recomendador Colaborativo . . . . .   | 14 |
| 2.6. Sistema Recomendador basado en contenido (Content-based)  | 17 |
| 2.7. Sistema Recomendador basado en el conocimiento . . . . .  | 21 |
| 2.8. Sistema recomendador en salud . . . . .   | 29 |
| 2.9. Ejemplo de una red semántica . . . . .  | 30 |
| 2.10. Análisis del Sentimiento y los Sistemas Recomendadores en<br>Salud . . . . .                   | 33 |
|  |    |
| 3.1. Algoritmo de recopilación de vídeos desde Youtube . . . . .                                     | 38 |
| 3.2. Resultado web de Saluteca . . . . .   | 42 |
| 3.3. Detalle de un vídeo Saluteca . . . . .  | 43 |
| 3.4. Resultado web de Diavideos . . . . .  | 46 |
| 3.5. Incorporación automática de contenidos en el portal . . . . .                                   | 48 |
| 3.6. Incorporación automática de contenidos en el portal . . . . .                                   | 49 |
| 3.7. Algoritmo de recomendación propuesto . . . . .  | 55 |
| 3.8. Resultados obtenidos de la etapa 3.3.2 . . . . .  | 56 |
| 3.9. Algoritmo de recomendación HealthRecSys de la seccion 3.3.3                                     | 58 |
| 3.10. Ejemplo de modelos CAS - fuente: cTAKES . . . . .  | 62 |
| 3.11. Nuevo modelo de HealthRecSys de la sección 3.3.4 . . . . .                                     | 63 |
| 3.12. Demo de la encuesta . . . . .  | 65 |
| 3.13. Resultados en CBMS . . . . .   | 69 |

|   |    |
|---|----|
| 3.14. Resultados en Diabetes . . . . .                  | 70 |
| 3.15. Resultados en Hipertensión/HTA . . . . .          | 71 |
| 3.16. Encuesta modelo de Precision@k Diabetes . . . . . | 72 |

# Índice de tablas

|  |    |
|--|----|
| 2.1. Matriz de interacción Usuario/Ítem . . . . .                  | 13 |
| 2.2. Ítem recomendados por el usuario X . . . . .                  | 13 |
| 2.3. Matriz de ítems con sus propiedades . . . . .                 | 18 |
| 2.4. Matriz Verdad/Falso de transacciones por cada ítem . . . . .  | 23 |
| 3.1. Entidad HealthVideo . . . . .                                 | 44 |
| 3.2. Entidad HealthChannel . . . . .                               | 45 |
| 3.3. Entidad HealthPlaylist . . . . .                              | 45 |
| 3.4. Entidad Estado HealthStateDrupal . . . . .                    | 46 |
| 3.5. Entidad Estado HealthStateDrupal . . . . .                    | 47 |
| 3.6. Valoración de coeficiente Kappa . . . . .                     | 50 |
| 3.7. Resultados análisis Algoritmo HealthTrust de la sección 3.3.3 | 60 |
| 3.8. Cálculo de las medias de Precision@k de enlaces recomendados  | 67 |
| 3.9. Cálculo de las medias de nDCG de enlaces recomendados .       | 68 |



# Lista de algoritmos

|    |   |    |
|----|---|----|
| 1. | Recuperación de videos de salud de Youtube™ . . . . . | 47 |
| 2. | Algoritmo recomendador . . . . .                      | 59 |
| 3. | Función getBioOntologyData . . . . .                  | 59 |
| 4. | función nDCG(x) . . . . .                             | 68 |



# Lista de Fuente XML

|   |    |
|---|----|
| 3.1. Sección fichero configuración cTAKES . . . . . | 65 |
|---|----|





# Prefacio

Esta investigación se llevó a cabo durante los años 2012-2017 en el grupo de investigación de Robótica y Tecnología de Computadores del Departamento de Arquitectura y Tecnologías de Computadores en la Universidad de Sevilla. Durante el mes de Agosto 2012 y entre Septiembre a Diciembre del 2013 estuve de visitante en Northern Research Institute (NORUT) y The Research Council of Norway a través de la Universidad de Tromsø (UIT), en la segunda estancia fui asignado al proyecto Yggdrasil. Ésta es una beca doctoral de investigación que forma parte del proyecto Health-Trust del Tromsø Telemedicine Laboratory cofinanciado por el Norwegian Research Council.

Este trabajo no hubiera sido posible sin la ayuda, aliento y apoyo de muchas personas. Por este medio deseo agradecer a todas ellas que participaron en los trabajos que hicieron posible esta tesis.

Quiero agradecer a mis directores Jose Luis Sevillano Ramos y Luis Fernández Luque por todo su tiempo, apoyo incondicional y la retroalimentación. Han sido un apoyo profesional indispensable para guiar la investigación y los responsables de la creación de un entorno adecuado para poder sacar el reto de la tesis adelante, que no se hubiera completado sin la dirección competente de ambos.

Durante el proyecto Yggdrasil mobility programme 2013-2014 (id. 227490/F11) tuve las pautas y orientaciones en la consecución y cumplimiento de objetivos, todo ello gracias a la coordinadora del proyecto Randi Karlsen de la Universidad de Tromsø. Esta beca se inició el 1 de Septiembre del 2013 hasta el 31 de Enero del 2014 por el que tuvo su aprovechamiento con la implementación del portal Diavideos y el inicio de recomendaciones basada en términos sujetos a las ontologías sanitarias.

El apoyo inestimable del profesor Gunnar Hartvigsen de la Universidad de Tromsø, los investigadores Ellen Brox, Lars Vognild, Stathis Konstantinidis y en general todo el Northern Research Institute (NORUT). Debo destacar las charlas y toda la atención del profesor Gunnar Hartvigsen. Todos pusieron ese grano de arena necesario y que brindó significado a la presente tesis.

Los fondos para la investigación fueron proporcionados en parte por la Universidad de Sevilla, Northern Research Institute (NORUT), The Research Council of Norway, Qatar Computing Research Institute (HBKU), Qatar Foundation y otra parte de forma personal. Deseo dar las gracias a todas las organizaciones por su apoyo. Uno de los principales problemas que se encontraron en la elaboración de esta tesis fue no poseer dedicación completa como es habitual en otras tesis y la falta de suficiente cobertura económica, aunque esto no ha impedido poder llegar a su finalización.

Un especial agradecimiento a mi familia por su amor, apoyo, compañía y paciencia. Aunque la realización de la investigación ha sido una gran parte de mi vida durante los últimos años, ésta no tendría sentido sin su ayuda y comprensión, gracias.

No quiero dejar de mencionar especialmente a mi madre por su amor, apoyo y ternura a lo largo de los años. Además quiero dedicar esta tesis también mi padre, no dudo que estaría dándome esos bocados de costumbre cuando aprobaba las asignaturas. También quiero dedicar un agradecimiento al Dr. Jose Antonio Rosell Antón por el que sin duda ayudó a mi madre en decisiones tan claves en mi infancia.

También quiero mencionar la labor mentora llevada a cabo por Alejandro Jadad, Julio Lorca y Lucia Quiroga, influyeron mucho en mi decisiones.

Además, estoy agradecido a mis amigos e investigadores, que han contribuido con su ayuda, apoyo y paciencia a esta tesis, indirecta o directamente. Especialmente a Rodrigo Román, Esteban Espinilla, Javier Pretel, María Pretel, Alejandro Rivero, Carlos Rizo, Francesc Saigí, Carlos Fernández Llatas, Ignacio Basagoiti, Manuel Armayones, Paloma Peláez, Rafael Pastor, Alejandro Miranda y José Manuel Meza.

A mis compañeros de trabajo especialmente a Francisco J. Alcázar, Carlos Pérez, Juan Carlos Durán, Francisco Acevedo, Enrique Rando, Francisco Carrión, Consuelo Egea, Francisco J. Guerrero, Enrique Orellana, Javier Escaño, Pablo Ávila, Alberto Amo, Rafael Márquez, Juan Antonio Reina,

Trinidad Pastor y Luis Manuel Pérez.

No quiero excluir a todos los profesionales sanitarios que han contribuido de manera desinteresada y voluntaria en la elaboración de encuestas -aguantando test enormes y que lo han atendido con toda la profesionalidad y seriedad que merece-. Entre todos ellos debo destacar a Francisco Sánchez Laguna, Pedro García Fortea, Gonzalo Aranda, Amalia Arce, Margaret Hansen, Xose Meijone, Alejandro Luque, Elena Sainz, Javier González de Dios, Hugo Briceño, Humberto Fernan, Marta Cano y Clara Traverso.

Por último, mi más sincero agradecimiento a Pablo Ávila, Salvador Portillo, Sergio Ruiz y Pablo Sánchez por ayudarme a mejorar en el inglés y ayudarme en las traducciones.

Gracias por siempre en darme el apoyo, la libertad y la comprensión en todas mis decisiones y elecciones.

Sevilla 25, de Abril de 2017.

Carlos Luis Sánchez Bocanegra.



# Preface

This research was carried out between the years 2012 and 2017 within the Computer Technology and Robotics research group of the Computer Technology and Architecture Department of the University of Seville. During the month of August 2012 and between September and December 2013 I was a visiting Researcher in Northern Research Institute (NORUT) and The Research Council of Norway through the University of Tromsø (UIT). During the latter traineeship I was assigned to the Yggdrasil project. This was a research fellowship that is a part of the HealthTrust project of Tromsø Telemedicine Laboratory cofinanced by the Norwegian Research Council.

This work would not have been possible without the help and support of many people. Therefore I would like to thank all those persons that took part in the works that made possible this thesis.

I would like to thank my thesis directors Jose Luis Sevillano Ramos and Luis Fernández Luque for all their time, unconditional support and the feedback. They were an invaluable professional support for guiding the research and the responsables for creating an appropriate environment for successfully carrying out the challenge of the thesis; it would have not been finished without their competent direction.

I would like to thanks specially to Randi Karlsen, the project coordinator in the University of Tromsø who gave me advice for accomplishing the objectives of the project during the Yggdrasil project mobility programme 2013-2014 (id. 227490/F11) This fellowship began the 1st of September of 2013 and ended the 31 of January 2014 . The results of this stay were the implementation of the Diavideos portal and the starting point for recommendations based in terms of linked to health ontologies.

The invaluable support of Professor Gunnar Hartvigsen of the Univer-

sity of Tromsø, the researchers Ellen Brox, Lars Vognild and of the whole Northern Research Institute (NORUT). I should highlight the discussions and all the attention of Professor Gunnar Hartvigsen. All of them contributed to this work and gave meaning to the present thesis.

The funding for the research were provided by the University of Seville, Northern Research Institute (NORUT), The Research Council of Norway, Qatar Computing Research Institute (HBKU), Qatar Foundation and also by personal funds. I would like to thank all the organizations for their support. One of the main problems that they found in the build-up of this thesis was not having full time dedication as it is common in other instances and the lack of enough economic support, though this did not prevent its conclusion.

I would specially like to acknowledge my family for their love, support, company and patience. Though the accomplishment of the research has been a big part of my life during the last years, it has not any meaning without their support and understanding, thanks.

I would like to mention and thanks to my mother for her love, support and tenderness along the years. Furthermore, I would like to dedicate this thesis to my father, I do not hesitate that he would be biting me as he usually did when I passed the subjects. I also want to acknowledge to Dr. José Antonio Rosell Antón who, with no doubts, helped my mother in important decisions during my childhood.

Thanks to Alejandro Jadad, Julio Lorca and Lucía Quiroga for the mentoring work carried out. They highly influenced in my professional scope.

Moreover, I am thankful to my colleagues and research fellows, that contributed with their help, support and patience to this thesis, directly or indirectly. Especially to Rodrigo Román, Esteban Espinilla, Javier Pretel, María Pretel, Alejandro Rivero, Carlos Rizo, Francesc Saigí, Carlos Fernández Llatas, Ignacio Basagoiti, Manuel Armayones, Paloma Peláez, Rafael Pastor, Enrique Dorronzoro, Francisco Grajales, Iván Zambrana, Alejandro Miranda and José Manuel Meza.

All of my friends and coworkers as Francisco J. Alcázar Martín, Carlos Pérez Alonso, Juan Carlos Durán, Francisco Acevedo, Enrique Rando, Francisco Carrión, Consuelo Egea, Enrique Orellana, Javier Escaño, Francisco J. Guerrero, Alberto Amo, Rafael Márquez, Juan Antonio Reina, Trinidad Pastor y Luis Manuel Pérez.

I would not like to forget all the health professionals that contributed voluntarily in the making of polls - going through huge tests that they covered with the professionalism and seriousness that the matter deserves. Among them I would like to enhance the role of Francisco Sánchez Laguna, Pedro García Fortea, Gonzalo Aranda, Amalia Arce, Margaret Hansen, Xose Meijone, Alejandro Luque, Elena Sainz, Javier González de Dios, Hugo Briceño, Humberto Fernan Mandirola, Marta Cano and Clara Traverso.

Finally, my deep and sincere gratitude to Pablo Ávila, Salvador Portillo, Sergio Ruiz and Pablo Sánchez Gámez for helping me improving my English and the translations.

I am eternally thankful for providing the support, freedom and understanding in all my decision and choices.

Seville Abril 25th 2017

Carlos Luis Sánchez Bocanegra.





# Capítulo 1

## Introduction

Health promotion is one of the main tasks of Public Health [McKenzie et al., 2016]. Traditional channels are used to deliver messages such as: advertisements, scientific journals as well as meetings or awareness talks. All of them are focused on increasing awareness of the importance of health care and inviting healthcare professionals to interact with their patients.

However these mechanisms have shown a series of specific problems linked to traditional communication channels such as: cost of the ads, the fact that the focus of the magazine is more oriented to professionals or the impossibility of elaborating long and clear messages on meetings or talks. In addition the rate of patients per doctor is in constant growth, reducing the care ratio as well as the quality during intervention time [Wilson et al., 2016].

Patients demand information to be able to control and manage their illnesses. How is this new challenge addressed in today's society? The birth of the first website in 1990 resulted in a paradigm change. Internet allows anyone to dump information in a quick and inexpensive fashion. This allowed to open a new framework of communication that revitalized the publication of health information, adding content in different formats, providing information and integrating it into the environment [McMullan, 2006].

A growing percentage of the population, especially in Western countries, have access to the internet and social media on a daily basis. The profile of these users used to be young IT savvy people. However today cost and access barriers have lowered and the use of navigation tools and

formats has been simplified. This has facilitated the provision of a wide range of content accessible to people of all age groups through the network [Vance et al., 2009]. In the field of health, internet provides a path for exchange of extremely useful information not only for patients, their families and caregivers, but also for anyone interested in their health condition.

However new challenges have also emerged. These are related to the proliferation of content which sometimes is not related to health interests, which can be harmful. This type of information is stranded with that published by reliable sources, therefore creating a health information system that is often confusing and possibly harmful for patients. This situation is forcing health professionals to increase surveillance and to become a sort of validated information guides. In this respect, the role of search and recommendation engines may be crucial to support, simplify and facilitate the complex task of the experts or health professionals by separating and selecting valid information from that which is not.

## 1.1. Approach

When we talk about health information on the internet we immediately think of information collected by professionals and others working collaboratively to compile it. This is the most orthodox and reliable source. However the behavior of patients and their relatives in relation to the care of their disease has been in continuous transformation since the birth of the internet. With a more proactive attitude, users tend to search for information related to their health on the internet, finding alternative sources to those used by health professionals. These alternative sources gather a huge volume of data which may include incorrect or outdated information not informed by the latest scientific advances. . . . Anyone can even find content generated different to their own health goals. For example, in pro-anorexia and anti-vaccines communities, etc. According to the case, the access to these alternative sources of information may be counterproductive or even pernicious.

As we will discuss in this thesis, one of the major emerging problems in the medical community is precisely how to manage health information generated on the Internet. The communication tools discussed above are different from the traditional ones, which derives in a change in the health

care model as we see it today. This new point of view requires health professionals to improve their observations and become reliable guides in order to provide access to curated and validated information.

The task of generating health content from various sources, as well as the control and selection of reliable information requires the formalization and validation of health contents on the internet. Therefore, it is necessary to approach this task from a formal and rigorous point of view. For this purpose an essential concept is that of “Ontology”, which can be defined as a specific vocabulary of representation that consists of definitions of classes, relations, functions, and other objects in a specific domain. Modeling contents based on ontologies, defining the most appropriate content, their preferences or features for a given patient could be identified on the basis of his/her disease.

This thesis is based in a so-called recommendation in health systems. In particular on one based on content, item or articles with health information sought by patients or their family members. We will study how these systems analyze the content that it recommends, how to recover it and treat it taking into account the heterogeneity of sources and formats. In particular, we will use ontologies as health dictionaries with additional information. Based on natural language processing techniques, we can interpret a document at different levels, and automate its morphological, syntactic and semantic understanding. In summary, the use of health ontologies, together with natural language processing, will allow to produce a personalized recommendation of health content, which will be evaluated experimentally.

## 1.2. Thesis structure

The structure of this thesis is following:

- 1st chapter: General introduction to the thesis, defining the objectives.
  - 1.1st section: Describes the motivation for the development of the thesis, focusing on the approach on health recommendations.
  - 1.2nd section: Describes the content distribution.
  - 1.3rd section: Objectives of the thesis.

- 1.4th section: List of projects and articles that emerged or are directly related to the thesis.
- 2nd chapter: Review and study of the State of the art. B apendix finds the english versión from chapter 9 from Methods in Molecular Biology (Clifton, N.J.). 1246: 131-46. 2015.
- 3rd chapter: Central chapter of the thesis, describes materials, research methods used and results focusing on the contributions and progress: content analysis, recommendations collection and its corresponding evaluation.
- 4th chapter: Critical discussion of the work and future lines of research are presented.
- A apendix: Setup code corresponding to the different chapter 3 contributions.
- B apendix: Chapter 9: Introduction on Health Recommender Systems from Methods in Molecular Biology (Clifton, N.J.). 1246: 131-46. 2015.
- C apendix: HealthRecSys: A semantic content-based recommender system to complement health videos Paper for BMC Medical Informatics and Decision Making.

### 1.3. Objectives

1. To analyze and obtain access channels and resources for patients in Web 2.0 environments.
2. Search repositories with sources of health resources, to share, validate and manage existing ones and create awareness among patients.
3. Design and implement recommendations algorithms based on collaborative techniques.
  - a) Algorithm to collect videos from a platform.
  - b) Algorithm Advisor in health.

4. Recommended contents that are positively valued by users with similar characteristics. For example, a pregnant woman with diabetes will receive recommendations to contents that have been highly valued by other women with similar problems, or recommended by professionals for people in a similar situation.
5. Usage of social network analysis and natural language processing techniques to estimate characteristics such as reputation or popularity.
6. Evaluation of recommender algorithm.

## 1.4. List of projects and publications

which are referred to in the text by their Roman numerals.

The development of this thesis has produced several publications and projects referred in the text by their Roman numerals.

- I Chapters 1.2: “New Technologies, New Actors and Chapter 1.5: Reflections on the E-Patient White Paper“. Salcedo, V. T. The ePaciente and social networks. ITACA-TSB. 2011. ISBN: 978-84-694-4110-7. 2011.
- II Sánchez-Bocanegra CL, Rivero A, Fernandez-Luque L, Sevillano JL. Saluteca, a Spanish Health Video Portal. Proceedings IV Workshop on Technology for Healthcare and Healthy Lifestyle. 2012.
- III Sánchez-Bocanegra CL, Rivero-Rodriguez A, Fernández-Luque L, Sevillano JL. Diavideos: the diabetes health video portal. Studies in Health Technol Informatics. 192: 941. 2013.
- IV Rivero-Rodriguez A, Konstantinidis ST, Sanchez-Bocanegra CL, Fernandez-Luque L. A health information recommender system: Enriching YouTube health videos with Medline Plus information by using SnomedCT terms. IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS). P. 257-61. 2013.
- V Chapters 3.5: ”Health Apps“ and Chapter 4.2: “ Reliability of Information ”. Basagoiti, I., Almendro, C., Armayones, M., Avila, J., Baci-

- galupe, G., Borile, M., ... Díaz, H. Health Literacy: From Information to Action . ITACA-TSB. ISBN: 978-84-695-5267-4. 2014.
- VI Yggdrassil Doctoral Researcher Grant: The project is focussing on the study of Recommender System technologies within the diabetes domain. Norwegian Research Council. 2014.
- VII Sanchez-Bocanegra CL, Sanchez-Laguna F, Sevillano JL. Introduction on health recommender systems. *Methods in Molecular Biology* (Clifton, N.J.). 1246: 131-46. 2015.
- VIII Fernandez-Luque L., Sánchez-Bocanegra C., Ramos J.L, Karlsten R. A Semantic Web Health Recommender System: Enriching YouTube Health Videos. *Medicine 2.0 Congress* available at <http://www.medicine20congress.com/ocs/index.php/med/med2014b/paper/view/2455>. 2015.
- IX Sánchez-Bocanegra C., Sevillano J.L., Rizo C, Civit A., Fernandez-Luque L. HealthRecSys: A semantic content-based recommender system to complement health videos. *Medical Informatics Decision Making - BioMed Central Journal*. DOI: 10.1186 / s12911-017-0431-7. 2017.

## Capítulo 2

# Estado del arte

Los sistemas de información y comunicación (TIC) proveen de distintos mecanismos de búsqueda y recolección de información relativa a la salud. Los internautas realizan búsquedas que les permiten acceder a una gran cantidad de información de diferentes fuentes a través de la Red [Fox and Jones, 2011]. Surgen las figuras de los prosumidores -internautas generadores de contenido- que vuelcan información en Internet. Esto puede provocar una saturación de información que puede llegar a colapsar a internautas interesados. Ante tal colapso los motores de búsqueda deben encontrar medios y métodos para encontrar información de salud fiable.

Los usuarios están por tanto demandando información de salud precisa y válida. Los buscadores son los encargados de esta tarea y muchos estudios analizan cómo funcionan y escogen la información relevante [Fernandez-Luque et al., 2012]. Los internautas solicitan información -en nuestro caso, información de salud- acorde con su perfil y necesidades. De esta manera el citado *sistema recomendador de salud* debe encargarse de encontrar esa información y adaptarla al perfil del usuario, que puede recuperar a través de la historia clínica electrónica o el registro personal de salud [Wiljer et al., 2008, Cline and Haynes, 2001].

Los métodos de recomendación dependen principalmente del perfil de los usuarios, la información de salud a buscar (ítems) y el contexto que les rodea. Los sistemas recomendadores en salud han presentado diferentes dificultades y problemas para poder tener los resultados esperados [Adomavicius and Tuzhilin, 2005]: el problema del arranque frío (cold-start),

serendipia, esparcimiento de los datos, spam entre otros. Esto exige que los sistemas recomendadores requieran de mejoras para extraer elecciones adecuadas.

En este capítulo, haremos una breve introducción a los sistemas recomendadores, definiendo sus principales elementos y revisando las distintas alternativas y los más recientes avances sobre todo en el ámbito de los sistemas recomendadores en salud. La mayoría de este material fue publicado como capítulo de libro en [Sánchez-Bocanegra et al., 2015] y servirá de base para la descripción en los siguientes capítulos del sistema recomendador desarrollado en la presente Tesis.

## 2.1. Elementos partícipes en la recomendación

En un sistema recomendador vamos a contar con una serie de entidades que participan en el entorno de la recomendación. Procedamos a estudiar cada una de estas entidades.

### 2.1.1. Perfil de usuario

El *perfil de usuario* es la entidad que contempla todas aquellas propiedades asociadas al usuario que participa (ver figura 2.1 y 2.2 <sup>1</sup>). Por ejemplo, un paciente de 28 años, de Irlanda y que padece de Diabetes Tipo 1 con niveles en sangre de  $11.2 \text{ mmol/L}$ . El sistema recomendador recopila toda esa información (lugar y fecha de nacimiento, tipo de Diabetes, nivel de azúcar en sangre) y lo incluye en su banco de datos. Esta información facilita la personalización de la recomendación.

### 2.1.2. Ítems

La siguiente entidad son los *ítems*: elementos que un usuario busca de acuerdo a sus necesidades (ver figura 2.3 <sup>2</sup>). Los ítems se representan como documentos que contienen propiedades con información de interés para el usuario: control de medicación, dosificación, prospecto, vídeos que describen el padecimiento o la enfermedad, etc.

---

<sup>1</sup>fuelle - <http://vis.berkeley.edu/courses/cs160-sp09/wiki/index.php>

<sup>2</sup>fuelle Youtube<sup>TM</sup>



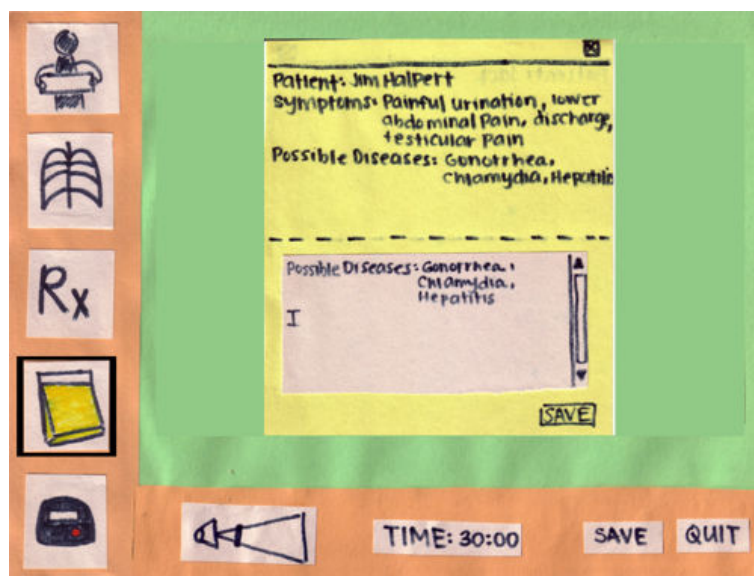


Figura 2.1: Esquema de un perfil - Fuente berkeley.edu

Todas esas propiedades se combinan y se incluyen en el proceso para ajustar la correcta elección. Los ítems representan la esencia de la recomendación [Jannach et al., 2010] ya que son los elementos que se ofrecen a los usuarios.

### 2.1.3. Contexto

La siguiente entidad es el *contexto*, que define aquellas propiedades externas que añaden información al sistema (ver figura 2.4 <sup>3</sup>), por ejemplo el posicionamiento georeferencial o el clima puede ser factores externos que pueden ayudar a acotar una mejor selección. El contexto tiene un gran potencial puesto que en la mayoría de los casos son propiedades que

<sup>3</sup>Interaction Designs Foundation - <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/context-aware-computing-context-awareness-context-aware-user-interfaces-and-implicit-interaction>

The screenshot shows a 'Medical Profile' window with a blue border. On the left side, there is a vertical toolbar with five orange icons: a caduceus, a ribcage, an 'Rx' symbol, a clipboard, and a computer monitor. The main content area is white and contains the following information:

**Medical Profile**

**Biodata**

Name: Jim Halpert  
 DOB: 10/1/78      Age: 30      Sex: Male  
 Height: 6'2"      Weight: 160 lbs      Blood Type: A+

**Demographics**

Address: 2341 Esperanza Avenue  
 City: Santa Clara      State: CA      Zip: 95054  
 Home: 408-406-9022      Work: 408-373-4954      Cell: 408-345-0678  
 Race: Caucasian      Religion: Agnostic  
 Marital Status: Engaged  
 Occupation: Assistant Regional Manager - Dunder Mifflin, Scranton

**Chief Complaint**

Symptoms: Facial redness, Visible veins  
 Site: Cheeks around forehead  
 Onset: Gradual  
 Character: Not painful, but rather inflammatory and irritating  
 Radiation: N/A  
 Associations: Facial flushing  
 Time Course: Appeared over the course of one year

At the bottom of the window, there is a status bar with 'Patient Status: Sick' and 'Time Remaining: 39:46'. On the right side of the status bar, there are two orange buttons labeled 'Save' and 'Quit'.

Figura 2.2: Ejemplo de un perfil de usuario - Fuente berkeley.edu

se pueden obtener automáticamente: por ejemplo la versión del sistema operativo, el navegador que utiliza, la hora de la consulta, tiempo que tarda en la elección de un ítem, etc.

#### 2.1.4. Dominio

El perfil de los usuarios, los ítems y el contexto son los principales elementos de un sistema recomendador. Interaccionan en el marco de lo que se denomina *dominio*. Por ejemplo pertenecen al dominio de Diabetes toda la información relativa a pacientes, investigadores, profesionales sanitarios y a todo el conjunto de información de diabetes. Los métodos recomendadores buscan la mayor aproximación con un dominio particular que pueda ser útil entre las entidades partícipes.

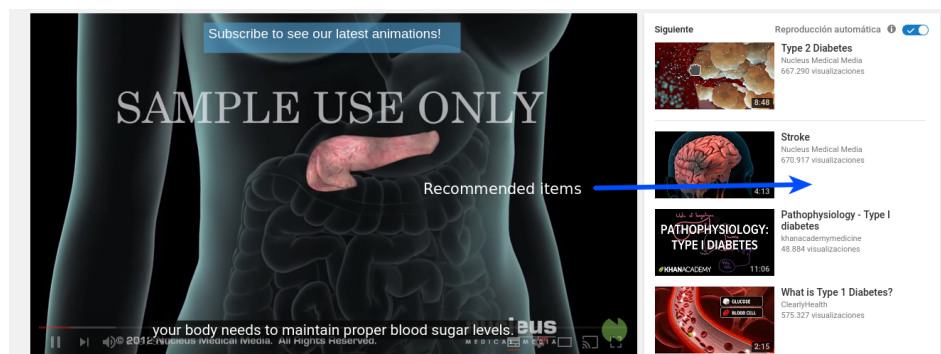


Figura 2.3: Ejemplo de recomendaciones (ítems) para Diabetes

## 2.2. Métodos básicos en los Sistemas Recomendadores

### 2.2.1. Colaborativos (Collaborative methods)

Uno de los principales métodos de recomendación son los sistemas colaborativos, enfocados al perfil de los usuarios y los ítems que han sido seleccionados en el pasado. Este método analiza el historial de selección en pacientes similares y es también conocido como método basado en el conocimiento (knowledge-based approach).

La tabla 2.1 representa por ejemplo este historial de selección almacenando en las filas a los usuarios y las columnas a los ítems. La intersección puede venir representado por valores enteros o booleanos. Esta tabla nos permite conocer el comportamiento y las similitudes o diferencias entre los usuarios con los ítems. También puede representarse como el formato que muestra en la tabla 2.2 por cada usuario.

La consideración básica en un sistema recomendador colaborativo es: *si en el pasado algunos usuarios compartieron el mismo interés por un ítem, probablemente en el futuro muestren por otro ítem similar el mismo interés* [Jannach et al., 2010] (ver figura 2.5). El método colaborativo responde a la siguiente cuestión: ¿Cómo puedo encontrar otros usuarios con intereses similares?, ¿Qué pasa con los nuevos usuarios? y ¿Qué pasa con los nuevos

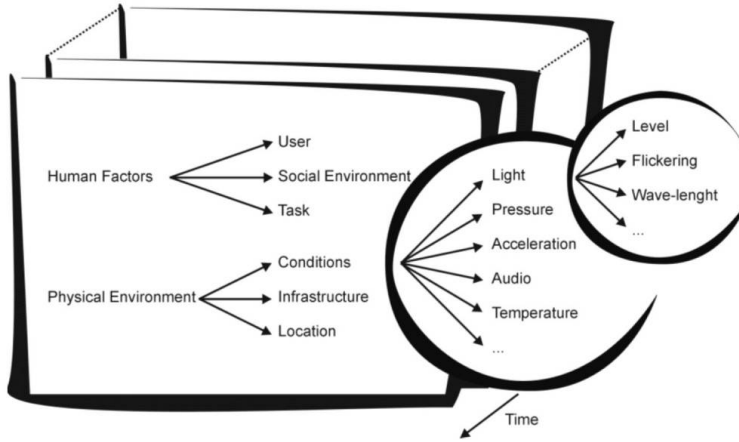


Figura 2.4: Estudio del contexto en una recomendación - fuente Interaction Designs Foundation

ítems?.

En este método, el rating de un ítem para un determinado usuario se calcula como la suma ponderada de los ratings de otros usuarios. El peso para cada término de esta suma ponderada representa la similitud entre los usuarios (a mayor parecido, mayor peso). Existen múltiples formas para obtenerlo [Broder and Ciccolo, 2004]. La correlación de Pearson [Benesty et al., 2009] es uno de los métodos utilizados para determinar el “vecino más próximo” (es decir, el usuario con mayor similitud). Aunque éste normalmente no ofrece la recomendación óptima por lo que sólo es válido en pequeños conjuntos de dominios. Este coeficiente se representa como  $\rho_{x,y}$  y su cálculo lo vemos en la fórmula 2.1, donde  $\sigma_{xy}$  es la covarianza “entre dos usuarios” (o entre los ratings de dos usuarios)  $x$  e  $y$ ,  $\sigma_x$  y  $\sigma_y$  representa la desviación típica.

$$\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (2.1)$$

Una alternativa es la medida de similitud de Cosine  $\cos(\theta)$  [Nguyen and Bai, 2011], donde los usuarios son considerados como vectores  $A$  y  $B$  en un espacio  $n$ -dimensional (siendo  $n$  el número de ítems). Así definimos una métrica

## 2.2. MÉTODOS BÁSICOS EN LOS SISTEMAS RECOMENDADORES 13

| <i>User/ítem</i> | <i>Ítem 1</i> | <i>Ítem 2</i> | ...   | <i>Ítem N</i> |
|------------------|---------------|---------------|---|---------------|
| <i>User 1</i>    |               |               |   |               |
| <i>User 2</i>    |               |               |   |               |
| ...              |               |               | <ul style="list-style-type: none"> <li>▪ Integer value</li> <li>▪ like/dislike</li> </ul> |               |
| <i>User M</i>    |               |               |   |               |

Tabla 2.1: Matriz de interacción Usuario/Ítem

| <i>Ítems recomendados</i> | <i>Ítem 1</i> | <i>Ítem 2</i> | ... | <i>Ítem Z</i> |
|---------------------------|---------------|---------------|-----|---------------|
|                           |               |               |     |               |

Tabla 2.2: Ítem recomendados por el usuario X

estándar que mide el ángulo entre los vectores. El resultado puede variar entre 0 (vector ortogonal, que significa que no hay similitud) y 1 (idénticos) conforme a la fórmula 2.2 [Adomavicius and Tuzhilin, 2005].

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (2.2)$$

Ambos métodos, correlación de Pearson y similitud de Cosine, pueden ser mejorados si restamos la media de ratings de los comportamientos o reacciones del usuario. De esta manera se toma en consideración el hecho de que diferentes usuarios puedan evaluar de manera diferente. El método resultante se llama medida ajustada de Cosine [Sarwar et al., 2001]. Otras opciones incluyen el coeficiente de correlación de Spearman [Myers and Sirois, 2014] conforme a la fórmula 2.3. Este coeficiente  $\rho$  determina la correlación entre dos variables aleatorias continuas  $x$  e  $y$ , con una interpretación similar al de Pearson. Para calcularlo, los datos son ordenados y reemplazados por su respectivo orden. Se genera entonces una nueva variable aleatoria  $D$  que es la diferencia entre los correspondientes nuevos valores ( $x - y$ ). Si  $N$  es el número de parejas, el coeficiente queda como en la referenciada fórmula 2.3.

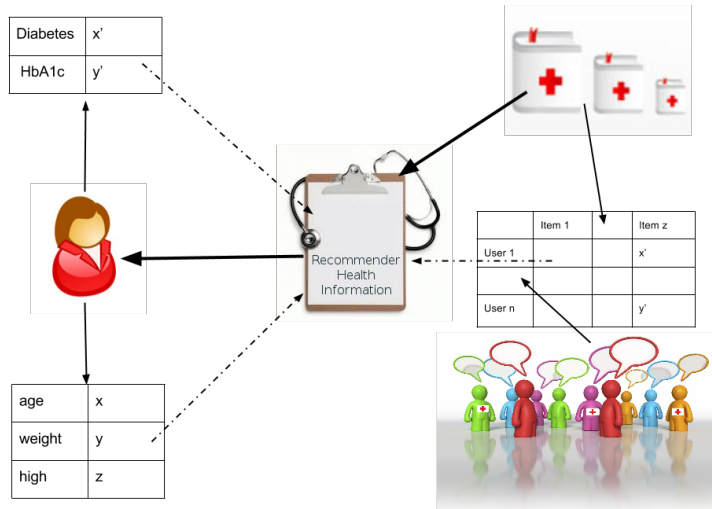


Figura 2.5: Sistema Recomendador Colaborativo

$$\rho = 1 - \frac{6 \cdot \sum(D^2)}{N \cdot (N^2 - 1)} \quad (2.3)$$

Otra soluciones más elaboradas que pueden encontrarse [Broder and Ciccolo, 2004]:

- *Frecuencia inversa del usuario (Inverse user frequency)*: Reduce la importancia relativa de aquellos casos que reciben de forma universal una misma valoración; tales casos reciben muchas valoraciones pero éstas no son demasiado útiles para discriminar. En la fórmula 2.4 podemos ver como se lleva a cabo su cálculo, donde  $n_i$  es el número de usuarios que han votado por el ítem  $i$  del total de  $n$  usuarios.

$$f_i = \log \frac{n}{n_i} \quad (2.4)$$

- *Significancia ponderada (Significance weighing)*: Dos usuarios que pueden tener gran coincidencia en gustos (altamente correlados) pero en

demasiados pocos ítems, en esos casos se puede utilizar la reducción de este método. Por ejemplo, exigiendo un mínimo de 50 rates, la predicción puede mejorar significativamente.

- *Amplificación de Casos (Case amplification)*: Dando más peso a los usuarios con mayor similitud incorporando un factor de aumento de valor en sus pesos. Para ello ese ajuste se aplica a los vecinos  $a$  y  $u$  con un factor multiplicador  $+1$  y  $-1$  de su peso original  $w_{a,u}$  y por una constante de factor  $\rho$  (ver fórmula 2.5 en el que  $\rho \geq 1$ ).

$$w'_{a,u} = w_{a,u} \cdot |w_{a,u}|^{\rho-1} \quad (2.5)$$

Un método diferente es tratar de reducir la complejidad computacional. Por ejemplo, cuando se selecciona el vecino más próximo con correlaciones positivas. Después de varias selecciones, este método tiende a los mismos ítems recomendados. Sin embargo, sufre de escasez o *sparsity* (insuficiente volumen de datos disponibles como para formular recomendaciones), y para evitarlo se puede utilizar otras alternativas como puede ser el método de los  $k$  vecinos más cercanos (K nearest neighbours -  $kNN$  algorithm) [Melville and Sindhvani, 2010]. En ambos casos la calidad de la recomendación se reduce.

Muchos de los métodos mencionados anteriormente se pueden adaptar para calcular las similitudes entre los ítems en lugar de los usuarios, y se ha demostrado que este enfoque puede proporcionar ventajas en términos de rendimiento computacional [Broder and Ciccolo, 2004].

### 2.2.2. Recomendación del vecino más próximo basada en ítem (Item-based nearest neighbor recommendation)

Define la similitud entre los ítems, normalmente acompañado de un procesamiento previo para poder deducir las recomendaciones subsiguientes en tiempo real en una fase “*online*” [Broder and Ciccolo, 2004]. Esta técnica suele identificar la similitud con proximidad calculando la distancia Euclídea conforme a la fórmula 2.6, donde  $p$  es el número de dimensiones de los ítems  $x_i$  y  $x_j$ , y  $x_{ri} - x_{rj}$  son la representación en la dimensión  $r$  del ítem  $i$ .

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^p (x_{ri} - x_{rj})^2} \quad (2.6)$$

Una mejor alternativa es usar las medidas Adjusted Cosine que consideran el comportamiento promedio de calificación del usuario [Symeonidis et al., 2008].

El procesamiento de la matriz entera de los usuarios y las relaciones con los ítems en tiempo real podría consumir una gran cantidad de tiempo y recursos. Para evitar eso, todos estos datos pueden ser procesados previamente para obtener una matriz de ratings por elemento. Este pre-procesamiento es un enfoque habitual en *máquinas de aprendizaje (machine learning)* [Tan, 2005]. Sin embargo, la obtención de esta matriz es muy difícil debido a que requiere un mínimo número de valoraciones del ítem por distintos usuarios así como un número limitado de vecinos [Linden et al., 2003].

Por último, vale la pena mencionar el método *Slope One*, que pre-computa la diferencia media entre las calificaciones de los diferentes usuarios sobre un ítem determinado [Wang and Ye, 2009]. Este enfoque es más fácil de implementar, utiliza la regresión lineal siguiendo la fórmula 2.7 donde  $b$  representa esta diferencia promedio entre la valoración de dos usuarios. El uso de un modelo de regresión simplificado con un sólo parámetro reduce la necesidad de almacenamiento y el coste computacional, y en cambio ofreciendo resultados equivalentes.

$$f(x) = x + b \quad (2.7)$$

### 2.2.3. Métodos basados en contenidos (Content-based methods)

Este enfoque tiene en cuenta las propiedades de los ítems y perfiles de usuario como la esencia de la recomendación (véase el gráfico 2.6). La idea es que si sabemos más sobre el ítem y/o el usuario la recomendación será más precisa, aun siendo el número de recomendaciones anteriores o el tamaño de la comunidad de usuarios pequeña. Muchos autores lo denominan enfoque basado en el modelo (model-based method) [Jannach et al., 2010].

Cada elemento puede ser representado por una lista de propiedades (véase la tabla 2.3), que contiene todas las características principales para



## 2.2. MÉTODOS BÁSICOS EN LOS SISTEMAS RECOMENDADORES 17

ser identificados.

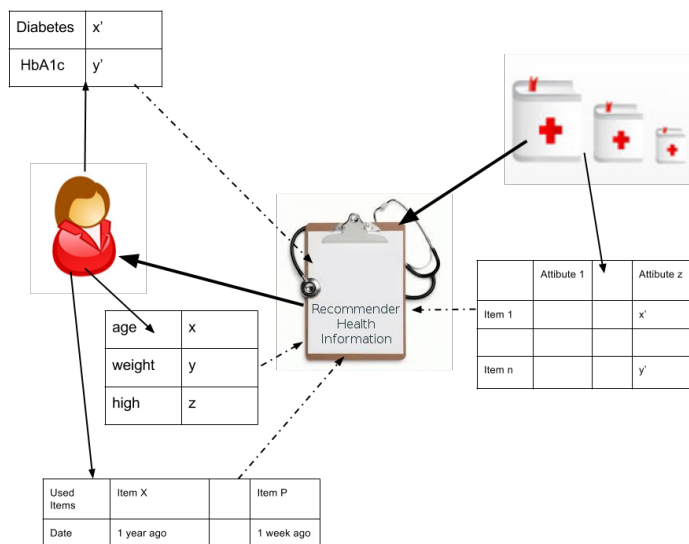


Figura 2.6: Sistema Recomendador basado en contenido (Content-based)

Primero necesitamos una fase previa de aprendizaje antes de realizar la recomendación (Offline). Sólo los modelos que han pasado por esta fase de aprendizaje se utilizan para hacer recomendaciones, por lo que el sistema necesita fases regulares de aprendizaje *continuo* con el fin de mejorar sus algoritmos y tomar mejores recomendaciones en cada iteración [Lops et al., 2011].

Este enfoque busca ítems similares a los que el usuario le podrían haber gustado en el pasado, encajando los atributos del ítem con la mejor selección acorde al perfil del usuario. A veces, las entradas en la tabla 2.3 no tienen que ser sólo metadatos, pueden ser palabras clave o términos asociados al ítem. En este caso, estos métodos se encargan de obtener *documentos sobre los ítems* (textos descriptivos y otros elementos) y filtrar si aparecen esas palabras clave o términos en dichos documentos. De manera similar a esta idea podemos encontrar los siguientes métodos básicos:

| <i>Item</i>   | <i>Tamaño</i> | <i>Peso</i> | <i>Precio</i> |
|---------------|---------------|-------------|---------------|
| <i>Item 1</i> | X             | Y           | Z             |
| <i>Item 2</i> | A             | B           | C             |
| ...           | ...           | ...         | ...           |
| <i>Item M</i> | X'            | Y'          | Z'            |

Tabla 2.3: Matriz de ítems con sus propiedades

### Modelo del vector del espacio vectorial TF-IDF (The vector space model and TF-IDF)

El método de espacio vectorial analiza todas las palabras clave o términos que aparecen en una clave o término y lo codifican como vectores con la dimensión del número de términos. Se utilizan dos medidas [Jannach et al., 2010]:

- *Frecuencia del Término (TF)*: Frecuencia con que aparece el término en el ítem. Existen muchas vertientes sobre el cálculo:
  - *Frecuencias booleanas*: Denotada como  $tf(t, d)$  representa un valor 1 si el término  $t$  ocurre en el documento  $d$  y 0 en otro caso.
  - *Frecuencia logarítmica*: Representada en la fórmula 2.8 donde  $f(t, d)$  es el número de veces que aparece el término  $t$  en el documento  $d$ .

$$tf(t, d) = \begin{cases} 1 + \log f(t, d) & \text{si } f(t, d) \neq 0 \\ 0 & \text{si } f(t, d) = 0 \end{cases} \quad (2.8)$$

- *frecuencia normalizada*: Representada en la fórmula 2.9.

$$tf(t, d) = \frac{f(t, d)}{\max f(w, d) : w \in d} \quad (2.9)$$

- *Frecuencia inversa de documento (IDF)*: una medida destinada a reducir el peso de los términos que aparecen con demasiada frecuencia y que no son muy útiles. Se representa en la fórmula 2.10 donde  $|D|$  es el número total de documentos y  $|d \in D : t \in d|$  representa el número de documentos donde aparece el término  $t$ .

$$idf(t, D) = \log \frac{|D|}{|d \in D : t \in d|} \quad (2.10)$$

El cálculo final de  $tf - idf$  se indica en la fórmula 2.11.

$$tf - idf(t, i, D) = tf(t, d) \times idf(t, D) \quad (2.11)$$

Téngase en cuenta que, aunque originalmente se utilizaba este método con los términos en los documentos, podría ser utilizado en otros contextos (por ejemplo, con películas si el contenido textual asociado está disponible). Hay varias mejoras de este método, principalmente en el sentido de reducir el tamaño de estos vectores y/o la cantidad necesaria de información, tales como evitar palabras inútiles o cortar el tamaño de la palabra, frase o contexto [Jannach et al., 2010].

El *método basado en la similitud del ítem (Similarity-based)* se basa en dos medidas principales:

- Similitudes de ítems.
- Likes/Dislikes en ítems anteriores

De esta forma, si un número dado de ítems similares gustaban antes, se recomienda un nuevo elemento. Utiliza el algoritmo del  $kNN$  y es simple de implementar [Suchal and Návrat, 2010], pero la precisión de la predicción es menor que la de los métodos más complejos.

El *método Rocchio* básicamente permite a un usuario evaluar los documentos, incorporando estas puntuaciones en el perfil del usuario. Esta retroalimentación se utiliza por el sistema para mejorar la consulta [Pazzani and Billsus, 2007]. El método funciona con *preguntas (queries)* representadas también como vectores. El método usa la fórmula 2.12, donde para una query  $\vec{Q}_i$  se tiene un conjunto de documentos relevantes y no relevantes asociados con pesos  $a$  (peso original),  $b$  (peso de los documentos relevantes) y  $c$  (peso de los no relevantes). El resultado es una query modificada  $\vec{Q}_m$  cuyo valor va en función de asociación de pesos de la query inicial  $\vec{Q}_0$  y los vectores de los documentos relevantes  $\vec{D}_j$  y no relevantes  $\vec{D}_k$ . Estos últimos términos se normalizan dividiéndolos por el tamaño

$|D_r|$  del conjunto de todos los documentos relevantes y  $|D_{nr}|$  en todos los documentos no relevantes.

$$\vec{Q}_m = (a \cdot \vec{Q}_o) + \left( b \cdot \frac{1}{|D_r|} \cdot \sum_{\vec{D}_j \in D_r} \vec{D}_j \right) - \left( c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\vec{D}_k \in D_{nr}} \vec{D}_k \right) \quad (2.12)$$

Los *métodos probabilísticos* utilizan un enfoque similar al de una tarea de clasificación, consistente en el etiquetado de los documentos de acuerdo con las calificaciones de usuarios anteriores. Algoritmos de clasificación simples como el Clasificador Bayesiano se han utilizado con éxito [Jannach et al., 2010, Barbieri and Manco, 2011].

El *aprendizaje automático (machine learning)* puede usarse para separar los términos relevantes y no relevantes. El modelo puede basarse en técnicas de aprendizaje automático como la agrupación, árboles de decisión, redes neuronales, etc. Un ejemplo es el algoritmo Widrow-Hoff/LMS (least-mean-square) que lleva a cabo el aprendizaje supervisado por el que el índice de rendimiento es el error cuadrático medio. Este algoritmo se compone de la red Adaline, (Adaptative Linear Neuron) con una función lineal de transferencia y la regla de aprendizaje LMS (Least Mean Square)[Lewis et al., 1996], ésta última es la que se aplica en el proceso de recomendación.

#### 2.2.4. Métodos basado en el conocimiento (Knowledge-based methods)

Estos métodos son un subconjunto del enfoque basado en contenidos descritos en la sección 2.2.3 (content-based). Se crea una base de conocimientos de las características del ítem con el fin de mejorar la recomendación (véase la Figura 2.7) solucionando algunas de las debilidades de las aproximaciones basadas en contenido 2.2.3. Por ejemplo, la recomendación basada en el conocimiento puede ser utilizada incluso con pocas calificaciones disponibles o útiles, ya que puede hacer recomendaciones apoyándose en mayor información sobre el usuario e ítem y no en las calificaciones explícitas anteriores.

El conocimiento acerca de los usuarios se puede obtener por ejemplo pidiéndoles sus preferencias y requisitos. Si los resultados del recomendador

## 2.2. MÉTODOS BÁSICOS EN LOS SISTEMAS RECOMENDADORES<sup>21</sup>

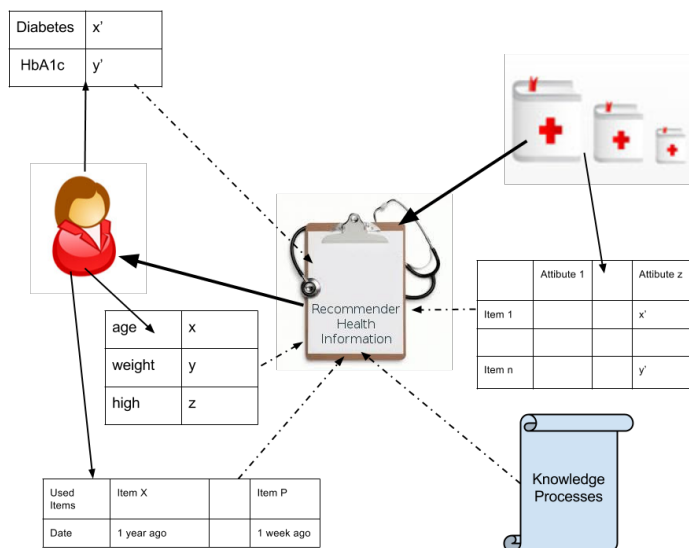


Figura 2.7: Sistema Recomendador basado en el conocimiento (Knowledge-based)

no son adecuados, el usuario puede pedir cambiarlos. Hay dos tipos básicos de recomendadores basados en el conocimiento:

- *Basados en casos (case-based)*: Tratan de encontrar otros ítems similares a los solicitados por el usuario. Para comprobar esta similitud se requiere de una descripción completa (conjunto de características) de los ítems [Bridge et al., 2005].
- *Basado en restricciones (restrict-based)*: Busca un conjunto de elementos que cumpla con las normas de recomendación. Este proceso es similar a la de un problema de satisfacción de restricciones o requisitos (Constraint Satisfaction Problem - CSP) en el que los requisitos y las características de artículos del usuario forman un conjunto de variables y el recomendador trata de obtener una solución para todas estas variables [Jannach et al., 2010, Yokoo et al., 1998].

## 2.3. Otros métodos de recomendación más avanzados

Los sistemas de recomendación se están aprovechando de algunos avances y mejoras en los métodos. A continuación describiremos brevemente algunos de los más relevantes.

### 2.3.1. Matriz de factorización/Modelos de factor latente (Matrix factorization/latent factor models)

Uno de los problemas comunes de la recuperación de la información consiste en encontrar los datos dentro de un gran volumen de documentos -algo particularmente importante para los métodos recomendadores. Dada una consulta de un usuario, una posible solución es encontrar datos que coincidan con la consulta. Sin embargo, este método tiene un conjunto de factores ocultos latentes (por ejemplo, sinónimos o palabras polisémicas en documentos textuales) [Jannach et al., 2010]. Las posibles soluciones incluyen:

- *Descomposición en valores singulares (Singular Value Decomposition - SVD)*: Reduce algebraicamente la dimensión de las matrices asociadas (formada por los usuarios e ítems). Una SVD de una matriz  $A$  es una factorización del tipo  $A = U\Sigma V^T$ , con  $U \in \mathbb{R}^{m \times m}$ ,  $V \in \mathbb{R}^{n \times n}$  ortogonales, donde  $m$  y  $n$  son ítems y usuarios, respectivamente.  $\Sigma \in \mathbb{R}^{m \times n}$  una matriz con los valores singulares de  $A$  en su diagonal principal ordenados de mayor a menor [Koren et al., 2009].
- *Análisis o Índice Semántico Latente (Latent Semantic Analysis/Indexing - LSA/LSI)*: descubre los factores latentes y reduce el tamaño de estas matrices mediante la fusión de los ítems semánticamente similares, a veces se usa en conjunción con SVD [Landauer et al., 1998, Hofmann, 1999].
- *Recuperación basada en LSI (LSI-based retrieval)*: permite recuperar los documentos correspondientes, incluso cuando no contengan muchas palabras de la consulta del usuario [Martínez-Torres, 2015, Dumais, 1991].

### 2.3.2. Minería de asociación de reglas (Association rule mining)

Este método encuentra patrones útiles en un conjunto de datos de transacciones. Las reglas de asociación están siguiendo la forma  $X \rightarrow Y$  donde  $X$  e  $Y$  son dos subconjuntos disjuntos de todos los elementos disponibles que satisfacen las restricciones sobre las medidas de la importancia y el grado de interés [Zhao, 2015, Hipp et al., 2000]. Este método es especialmente interesante en sistemas de recomendación de salud. Por ejemplo, supongamos que un usuario busca información sobre la Diabetes, si descarga un vídeo que explica las posibles complicaciones de la enfermedad un sistema de recomendación puede ofrecer vídeos sobre estilos de vida saludables.

Este método se representa en una matriz de interacciones de los ítems, cada celda de intersección contiene un valor booleano (ver Tabla 2.4):

- *True*: Significa presencia.
- *False*: Significa ausencia.

| #Transacción | Ítem 1 | Ítem 2 | ... | Ítem N |
|--------------|--------|--------|-----|--------|
| 1            | True   | False  | ... | True   |
| 2            | False  | True   | ... | False  |
| ...          | ...    | ...    | ... | ...    |
| n            | True   | False  | ... | True   |

Tabla 2.4: Matriz Verdad/Falso de transacciones por cada ítem

Cada matriz puede contener casi miles de transacciones. El método encuentra reglas que correlacionan la presencia de un conjunto de ítems con otro conjunto. Este método puede ser evaluado como válido si dados dos ítems  $x$  e  $y$ , encuentra con el soporte mínimo las reglas  $x$  e  $y \rightarrow z$  (probabilidad condicional de que una transacción con  $x$  e  $y$  contenga también  $z$ ) [Sarawagi et al., 1998].

### 2.3.3. Recomendaciones usando ontologías y semántica web

El Formato de Descripción de Recursos (RDF - Resource Description Format) es un estándar que representa la información de modelado como un *grafo*. RDF es uno de los pilares de la llamada Web Semántica, que puede utilizarse para serializar información representada mediante grafos. Junto con lenguajes como OWL (Ontology Web Language), estos elementos llevan a una mejor representación de la información que puede ser utilizada para obtener recomendaciones mejoradas [Agarwal et al., 2016, Akerkar et al., 2012, Capelle et al., 2012, Tomanek et al., 2007, Carroll et al., 2004].

Una ontología define un marco común consensuado que además de almacenar información permite buscar y recuperarla. Las ontologías pueden reforzar los sistemas de codificación tanto en los registros electrónicos de salud [Rector et al., 2009] como en las redes sociales [Bedi et al., 2007], reuniendo el significado preciso de un término, y la determinación de las relaciones entre los términos. Un término está hecho de palabras que representan un elemento. Cualquiera de los términos o las relaciones ayudan al sistema para mejorar las recomendaciones que hace [Suliman et al., 2016, Peis et al., 2008].

### 2.3.4. Métodos híbridos (Hybrid methods)

Los sistemas de recomendación híbridos combinan dos o más enfoques de recomendación para obtener un mejor rendimiento. Normalmente el filtrado de colaboración se combina con algunas otras técnicas (como basado en el contenido o conocimiento) con el fin de minimizar sus respectivas debilidades [Burke, 2002].

## 2.4. Desafíos existentes en los sistemas recomendadores

En esta sección describimos brevemente algunos de los desafíos a abordar en la investigación sobre los sistemas de recomendación.



### 2.4.1. Evaluación implícita y explícita (ratings)

Algunos sistemas recomendadores solicitan a los usuarios clasificaciones explícitas de ítems para obtener opiniones de usuarios más precisas (5 puntos, 7 puntos, like / dislike). El uso de estas calificaciones permite a estos sistemas usuarios dar recomendaciones más precisas a pesar de que requiere un esfuerzo adicional de los usuarios [Jannach et al., 2010]. Por ejemplo, sobre una escala de 10 puntos valorar una película.

Por otro lado, las calificaciones implícitas recogen propiedades externas del entorno. Por ejemplo, cuando alguien compra una lechuga en el mercado municipal de la ciudad de Málaga en verano a las 13 horas. Los parámetros del lugar y la hora se consideran calificaciones implícitas que pueden proporcionar recomendaciones saludables adecuadas para el clima, generalmente cálido y húmedo. En cuanto al estudio del comportamiento del usuario (comprar una lechuga) no se puede interpretar como correcta con tan pocas propiedades de su entorno.

La obtención de calificaciones útiles sin esfuerzos adicionales de los usuarios, complementado con el resto del dominio (como el contexto) es uno de los principales desafíos de los sistemas recomendadores.

### 2.4.2. Dispersión de los datos (data sparsity)

Otro problema típico de los sistemas de recomendación, particularmente importante con los que utilizan los métodos colaborativos, es que los datos y clasificaciones tienden a ser escasos contenidos en una matriz de calificación enorme. Una buena solución es utilizar la información de perfil de usuario en el cálculo de la similitud de otros usuarios [Adomavicius and Tuzhilin, 2005]. Es decir, si dos usuarios sufren de la misma enfermedad podrían considerarse similares, incluso si no han evaluado el mismo ítem de manera similar. Otras características como: género, edad, educación, intereses, podrían ayudar en la clasificación del usuario. Además métodos como los factores latentes o matriz de factorización se pueden utilizar para reducir el tamaño y dimensión de la matriz de calificación [Symeonidis, 2016, Žitnik and Zupan, 2015].

### 2.4.3. Problema del arranque en frío (cold-start)

Se necesita tiempo para incluir los nuevos usuarios e ítems. Puesto que a priori no se tiene conocimiento sobre ítems o usuarios, surgen varias preguntas: ¿Cómo se pueden hacer recomendaciones a los nuevos usuarios que no han calificado aún? Y ¿Cómo lidiar con los artículos que no se han votado aún?

Algunos enfoques evitan el problema de nuevos usuarios preguntando al usuario un número mínimo de calificaciones antes de que el servicio pueda ser utilizado. Otros explotan la supuesta 'transitividad' entre ellos a través de gráficos de relación [Jannach et al., 2010]. Pero el enfoque más habitual es la adopción de métodos híbridos, es decir, una combinación de métodos basados en métodos colaborativos y basados en contenido [Adomavicius and Tuzhilin, 2005].

### 2.4.4. Serendipia/Sobre-especialización (Serendipity/overspecialization)

Algunos recomendadores (principalmente basados en contenido) se apoyan en la similitud de los ítems y el interés del usuario por ítems similares se da por sentado. Sin embargo, en muchos contextos esta suposición es incorrecta. Por ejemplo, cuando el sistema sólo puede recomendar artículos que puntúan alto con respecto al perfil de un usuario. La recomendación puede ser demasiado obvia para ser útil. Este efecto indeseable se llama exceso de especialización o sobre-especialización. Por el contrario, aumentar la serendipia es generalmente útil. A veces el recomendador puede ofrecer artículos no buscados pero tal vez útiles, incluso con una cierta aleatoriedad [Zhang et al., 2012, Iaquinta et al., 2008].

### 2.4.5. Latencia

Este problema aparece en nuevos ítems (sobre todo con los métodos colaborativos), el sistema no es capaz de seleccionar los elementos añadidos recientemente que necesitan ser evaluados antes de que puedan ser recomendados [Sollenborn and Funk, 2002]. Esto es un inconveniente pues es deseable ofrecer las recomendaciones en un tiempo razonable.

## 2.5. Sistemas Recomendadores en Salud

Los usuarios de Internet buscan información de salud continuamente a través de las redes sociales [Choudhury et al., 2014], entre ellos llegan al 61 % en España, cifra que sigue en crecimiento [Marin-Torres Viviana et al., 2013]. Los internautas pueden identificar las razones de su enfermedad, encontrar tratamientos, aprender hábitos saludables o incluso poder contactar con otros pacientes con similares circunstancias o con profesionales de la salud. Ésto da como resultado un incremento contínuo de su conocimiento y *empoderamiento*.

Pero dado que el volumen de información es inmerso, muy a menudo resulta difícil de encontrar la información de salud realmente útil para el usuario. Poder identificar recursos válidos es una labor muy difícil -más aun obtenerla en un tiempo aceptable-. Por tanto, el rol del sistema recomendador ejerce una función vital para proporcionar la información deseada. El objetivo es buscar la fiabilidad, rendimiento y adaptación de los resultados de acuerdo al perfil y necesidad de los usuarios [Wiesner and Pfeifer, 2010].

Un concepto reciente relacionado con los sistemas recomendadores colaborativos es introducido por Gunther Eysenbach: *apomediaction*. Un término *socio-tecnológico* que identifica *una alternativa a la búsqueda de información y los servicios confiables y creíbles* [Eysenbach, 2008]. Frente a la búsqueda no guiada de información por parte de los propios usuarios, o la intervención de expertos que puedan proporcionar información fiable, una tercera vía consistiría en guiar a los usuarios hacia la información a través de unos agentes "apomediarios", que no son estrictamente necesarios para obtener la información pero que pueden mejorar la utilidad y fiabilidad de la misma. Pueden ser algoritmos (sistemas recomendadores) que filtran información de salud, encajándola a las necesidades del usuario.

Los sistemas colaborativos referidos en la sección 2.2.1 han sido los más utilizados en el sector salud [Broder and Ciccolo, 2004]. Sin embargo, las decisiones en salud dependen del conocimiento de las verdaderas necesidades del usuario. A veces ese conocimiento es incompleto generando unos resultados contrarios de lo que se busca.

### 2.5.1. Sistemas recomendadores y el registro de salud personal (PHR)

El registro de salud personal (PHR) es un registro que contiene toda la información acerca de los cuidados de un paciente [Nazi, 2013]. Pueden contener información demográfica, datos personales, de la persona de contacto en caso de emergencia, posiblemente incluyendo información sobre miembros de la familia. También mantienen información para cualquier tratamiento médico, cuidado, centros de salud, laboratorios y farmacias además de directrices e instrucciones de cómo atender a su salud.

Normalmente el registro de salud personal PHR lo gestionan los propios pacientes o cuidadores, ya sea en un dispositivo de almacenamiento o en la misma nube. Un enfoque útil consiste en tener un sistema de recomendación de salud que incluya la información de este registro PHR como parte del perfil de usuario [Wiesner and Pfeifer, 2010]. Por supuesto, hay muchos problemas de privacidad, de seguridad y éticos relacionados con este enfoque [Cushman et al., 2010]. En ese caso se plantea escoger valores del PHR menos sensibles para paliar este problema, omitiendo los datos que identifiquen a un paciente. El resultado final es que el perfil del usuario es considerado como parte del perfil mencionado en la sección 2.1.1.

En todos los métodos recomendadores (colaborativos, basado en contenido y basado en el conocimiento) puede asociarse el PHR al perfil del usuario en el conjunto de su dominio. De esta manera puede personalizar la búsqueda de una recomendación, siempre sujeto a cubrir los desafíos detallados en la sección 2.4. Podemos ver un ejemplo en la figura 2.8 donde el sistema recomendador puede ser de cualquiera de los tipos descritos anteriormente, o incluso una combinación de varios de estos tipos como en los modelos híbridos mencionados en la sección 2.3.4.

En los últimos años ha surgido un empuje creciente del uso de los dispositivos móviles para la salud [Hameed, 2003] lo que ha generado un nuevo modelo de flujo de información que facilita el acceso a información médica de una manera más rápida. Del mismo modo, los recomendadores se adaptan a esta nueva necesidad incorporando este modelo para facilitar y generalizar la óptima recomendación en estos nuevos formatos de presentación [Gavalas et al., 2014, Ricci, 2010]. Surgen conceptos como *UI (User Interface)* y *UX (User Experience)* para estudiar el comporta-

miento de los usuarios ante los distintos formatos y dispositivos con los que interactúa [Ahram et al., 2015], y de esta relación también se aprovechan los sistemas recomendadores [Park et al., 2015, Banos et al., 2015, von Niman et al., 2006].

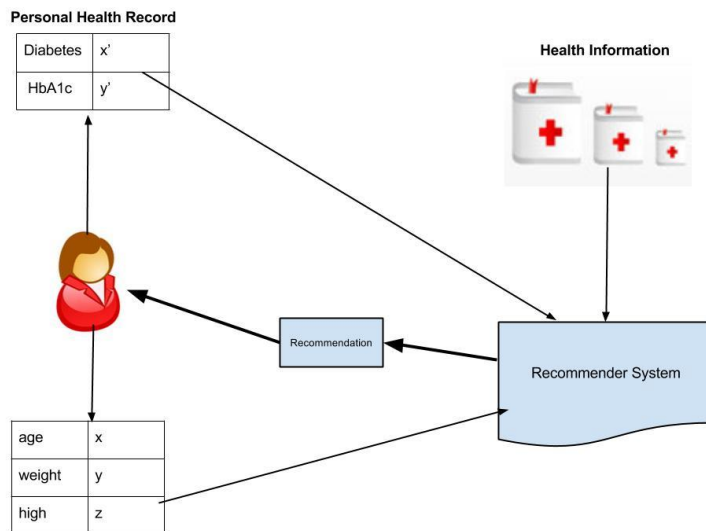


Figura 2.8: Sistema recomendador en salud

Una forma de integrar los sistemas recomendadores con el registro de salud personal PHR es a través de las redes semánticas, por ejemplo usando Wikipedia <sup>4</sup> [Wiesner and Pfeifer, 2010]. Una red semántica es la que representa relaciones entre conceptos. Cada concepto es una representación de conocimiento (también denominado como "sujeto"), y las relaciones (o 'verbos') y definen una acción entre dos conceptos. Normalmente se representan con un grafo en el que cada nodo es un concepto (sujeto) y los vértices son las relaciones (verbo) [Davoodi et al., 2013, Spitzer et al., 1993].

La figura 2.9 <sup>5</sup> muestra un ejemplo de red semántica en salud. Los nodos contienen conceptos de salud (sujetos) y los vértices asocian los conceptos

<sup>4</sup>Wikipedia - <http://www.wikipedia.org>

<sup>5</sup>Fuente JMIR - <http://www.jmir.org/2012/3/e73/>

con una acción/relación (verbos). Este ejemplo recogido de [Teodoro et al., 2012] incluye una breve representación de tablas en SQL. El sistema resultante utiliza la distancia entre ambos nodos como principal criterio de recomendación.

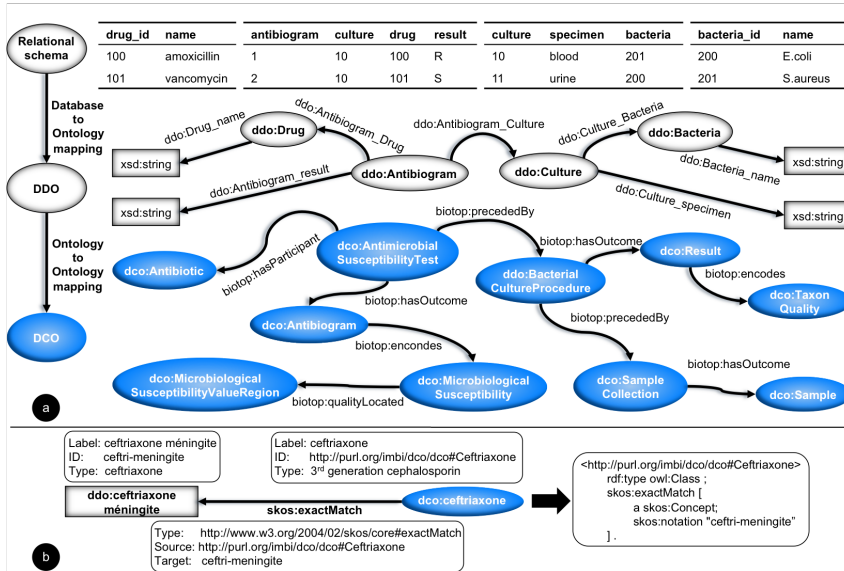


Figura 2.9: Ejemplo de una red semántica

## 2.5.2. Sistemas recomendadores y Social Media

El término Social Media hace referencia a un conjunto de recursos que habilita la gestión de contenido e información a través de Internet. Según [Kietzmann et al., 2011], los Social Media son *“plataformas sociales que crean y comparten contenidos generados por los usuarios”*. El eje central es el usuario en sí el cual forma parte integrándose en las redes sociales, foros, comunidades virtuales, etc.

Actualmente en Internet se vuelcan múltiples contenidos en un proceso que va en aumento. Las personas desean compartir sus experiencias y conocimientos a través de los formatos que Internet nos presta: blog, podcast,

vídeos, etc [Van de Belt et al., 2012]. El Social Media nos hace ser actores a la vez activos y pasivos en la generación de información en Internet. Esto puede ser beneficioso al poner en contacto personas con los mismos intereses, encontrando información, apoyo y consejos útiles. Pero también puede tener efectos no deseados como la difusión de malas prácticas o información engañosa.

Una forma de evitar estos efectos es aprovechar los recursos educativos de salud. La unión de los sistemas recomendadores y de estos recursos de alfabetización de salud [Basagoiti, 2012, Salcedo, 2011] nos permite filtrar la ingente cantidad de información disponible para poder ofrecer información relevante, centrada en las necesidades concretas de los usuarios. Podemos encontrar ejemplos donde se impulsa la recomendación de contenido personalizado a pacientes [Roitman et al., 2010a]. La información se acompaña con explicaciones detalladas para educar a los pacientes e incluso poder tomar decisiones médicas orientadas por un profesional sanitario.

### 2.5.3. Procesado de Lenguaje Natural

El NLP (Natural Language Processing) o Procesado de Lenguaje Natural es un área de la inteligencia artificial y lingüística que se enfoca en la comunicación entre el lenguaje máquina y el humano, usando técnicas y procedimientos que simulen la comunicación. Tiene como función extraer información sintáctica, semántica así como detalles cognitivos con incorporación de técnicas de aprendizaje automático [Chute et al., 2010, Kwon, 2000, Gao et al., 2015].

El aprendizaje automático (machine learning) desarrolla métodos que posibilitan un proceso de aprendizaje en las máquinas a partir de información estructurada, esta técnica se sustenta en el análisis de los datos con un fuerte componente estadístico. [Pang et al., 2002].

NLP debe atender a distintos niveles [Carbonell, 1994]:

- Nivel léxico: para determinar el significado correcto de un término que pueda tener distintas acepciones.
- Nivel referencial: Evitar repetición de términos y extraer la esencia en el contexto del contenido.

- Nivel estructural: Uso de la semántica que lleve a la construcción de árboles semánticos.
- Nivel pragmático: Frases fuera de contexto que no están sujetas a su propio significado sino por ejemplo a una ironía.

En función de los distintos niveles NLP gestiona los siguientes análisis:

- Análisis morfológico: estudia cada palabra de una oración en sí misma.
- Análisis sintáctico: estudia las reglas para la formación de unidades superiores como son las oraciones.
- Análisis semántico: estudia el significado en sí de las oraciones y palabras.
- Análisis pragmático: estudia el contexto que rodea a la frase, más allá de su significado semántico.

Dada la amplitud de posibilidades que brinda esta tecnología nos enfocamos en un punto que nos podría ayudar a mejorar la búsqueda de recomendaciones. En concreto, el análisis semántico sería el factor diferencial que nos permitirá mejorar las técnicas de recomendación planteadas anteriormente [Erekhinskaya et al., 2016].

### **Análisis del sentimiento - Machine Learning**

Machine Learning es una técnica de aprendizaje continuo, el sistema va mejorando al recoger cada experiencia o actuación. Se adapta y modifica en función de la interacción con los usuarios y entrega una decisión acorde a elecciones previas. Como hemos estudiado en el apartado anterior, esta técnica incrementa el potencial de los sistemas recomendadores [Harrington, 2012].

La técnica de Machine Learning es también muy eficiente en la clasificación de sentimientos de un usuario [Pang et al., 2002]. El análisis del sentimiento nos permite añadir más información subjetiva del contexto, por ejemplo el estado de ánimo, modo de expresión, etc. Los usuarios están constantemente comunicando sus sentimientos, así podemos ver el enorme



volumen de verbos que expresan sentimientos (sentir, querer, amar, etc) o sujetos (dolor, pérdida, sentimientos, etc). El análisis del sentimiento adapta la recomendación de acuerdo con el contexto sentimental del usuario, una aproximación básica la encontramos en la figura 2.10).

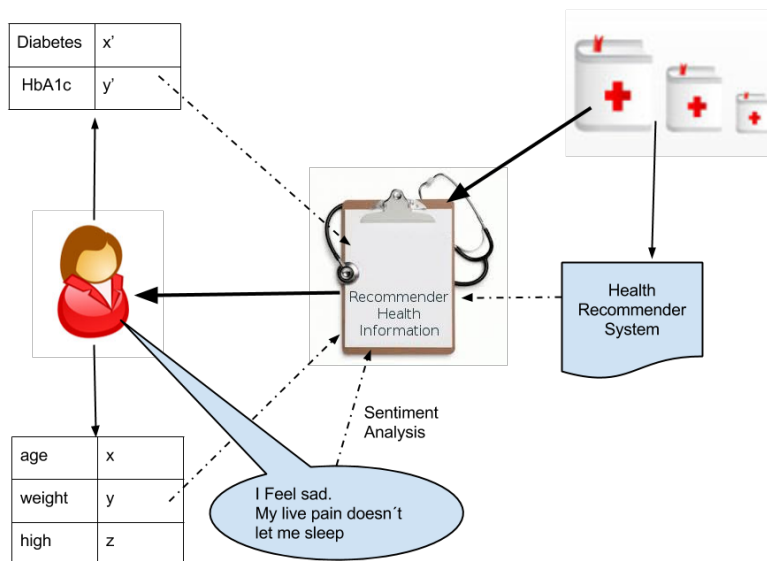


Figura 2.10: Análisis del Sentimiento y los Sistemas Recomendadores en Salud

La gestión de opiniones, sentimientos y contenido subjetivo es una de las ramas de investigación de mayor interés en la comunidad científica, especialmente en los sistemas recomendadores [Pang and Lee, 2008]. Los sentimientos nos proporcionan un elemento implícito que puede ser incluido en un sistema recomendador. Para ello podemos integrar tanto los medios sociales (Facebook<sup>TM6</sup>, Twitter<sup>TM7</sup>, Youtube<sup>TM8</sup>) como canales de

<sup>6</sup>Facebook <http://www.facebook.com>

<sup>7</sup>Twitter - <http://twitter.com>

<sup>8</sup>Youtube - <http://youtube.com>

dicados a la salud que incluyen ese tipo de mensajes (Patientlikeme<sup>TM9</sup>, Chrocane<sup>TM10</sup>, Webicina<sup>TM11</sup>). Toda esta información se puede recoger y procesar con fuentes válidas de información en salud, mejorando el funcionamiento de los sistemas recomendadores.

---

<sup>9</sup>Patient Like Me - <https://www.patientslikeme.com>

<sup>10</sup>Chrocane - <http://www.cochrane.org>

<sup>11</sup>Webicina - <https://www.webicina.com>

## Capítulo 3

# HealthRecSys: Recomendador en salud

### 3.1. Contexto de la aplicación

Las personas están cada vez más interesadas por tener el control de su salud. Este interés lleva tanto a pacientes como a cuidadores y familiares a buscar de forma activa información de salud relativa a su situación. A veces exigiendo un nivel de calidad que aumente las posibilidades del bienestar del paciente, otras veces con otros objetivos distintos que incluso pueden ser perjudiciales para su salud, como puede ser el caso de los colectivos anti-vacuna o pro-anorexia. A día de hoy nos encontramos con una excesiva -y a veces contradictoria- cantidad de información en la Red, lo que ha suscitado interés en la creación de servicios que suministren fuentes de información fiables. Uno de esos servicios son los sistemas recomendadores, que pueden usarse para filtrar soluciones adecuadas al paciente [Noveck, 2009].

Como se ha discutido en capítulos anteriores, los modernos sistemas recomendadores pueden llegar incluso a ofrecer información personalizada, útil y veraz.

Durante la estancia en el centro Northen Research Institute <sup>1</sup> se desarrolló un canal multimedia de contenidos orientados a la salud [Sánchez Bocanegra, 2011]. Como resultado de este proyecto surgió la idea de añadir mayor información

---

<sup>1</sup>Northen Research Institute - NORUT <http://norut.no/>

que apoye a cada vídeo. Dicha información permitiría dar recomendaciones sobre fuentes fiables de salud, como parte de un sistema recomendador en salud.

Para determinar qué tipo de sistema recomendador sería más adecuado, se analizó el estado del arte en los sistemas recomendadores -con especial hincapié en los sistemas recomendadores en salud-. Este estudio ha sido descrito en el capítulo 2, donde se discuten los distintos métodos y se observa cuáles son las últimas aportaciones, incluyendo sus particularidades: fortalezas y defectos.

En este capítulo vamos a explicar la realización de un sistema recomendador de vídeos de salud que denominamos *HealthRecSys*. Describiremos su estructura, y hablaremos del contexto de la aplicación. Explicaremos con especial detalle la forma en que llevamos a cabo la evaluación del sistema recomendador, que requirió la creación de dos portales, uno de salud en español y otro de Diabetes, para poder evaluar el prototipo: Saluteca [Sánchez-Bocanegra et al., 2012] y Diavideos [Sánchez-Bocanegra et al., 2013] son los portales que se crearon para tal fin.

### 3.1.1. Saluteca y Diavideos

Una de las plataformas para los llamados medios de comunicación (Social Media) mencionada en la sección 2.5.2 es Youtube<sup>TM</sup>, plataforma que permite ver y compartir contenidos de forma transparente, ágil y dinámica [Burgess et al., 2013] además de permitir la gestión (creación, actualización y eliminación) de vídeos. Desde 2005 es considerada la mayor comunidad de vídeos generados por los usuarios [Cheng et al., 2012, Cheng et al., 2008].

Los vídeos generan mayor impacto y alcance que los demás medios sociales. La facilidad para crearlos, editarlos y publicarlos en cualquier dispositivo hace que actualmente sean uno de los principales contenidos en los Medios Sociales <sup>2</sup> <sup>3</sup>. En el ámbito de la salud, más de 76 % de los sistemas de información hospitalaria incluyen vídeos para la alfabetización

---

<sup>2</sup>Social Media and Young Adults from Pew Internet <http://www.pewinternet.org/Reports/2010/Social-Media-and-Young-Adults/Part-2/1-C-ell-phones.aspx> available at <http://www.webcitation.org/6Bj9AN7mu>

<sup>3</sup>Teens and Social Media from Pew Internet <http://www.pewinternet.org/Reports/2007/Teens-and-Social-Media/1-Summary-of-Findings/1-Summary-of-Findings.aspx> available at <http://www.webcitation.org/6Bj9Mrpk1>

en salud al paciente, los familiares y los propios profesionales sanitarios. Son éstos quienes crean sus propios espacios de comunicación como las páginas en Facebook<sup>TM</sup> y canales en Youtube<sup>TM</sup> entre otros [Ventola, 2014].

Los elementos principales de Youtube<sup>TM</sup> son los vídeos y los usuarios [Chou et al., 2011]. Cada vídeo contiene *metadatos* que son datos adheridos al vídeo como el título, las etiquetas (tags), las categorías, la descripción, o el geoposicionamiento entre otras. Un usuario puede ser una persona, una empresa o una organización. De esta manera la plataforma facilita a estos usuarios la posibilidad de gestionar vídeos (crear, modificar y eliminar). Por un lado se permite la organización de vídeos a través del *Playlist* por el que se habilita la agrupación de vídeos con características comunes, a través del *Canal*, que facilita poder crear sus propios contenedores de vídeos.

Existen más de 800 millones de usuarios únicos que acceden a Youtube<sup>TM</sup>. Actualmente, se suben 72 horas de vídeos cada minuto, cerca de 4 billones de horas de vídeos están viéndose cada mes y 100 millones de usuarios interactúan a través de esta plataforma <sup>4</sup>. Estos valores nos dan una idea de la funcionalidad del portal Youtube<sup>TM</sup> y del enorme volumen de información que alberga.

Además Youtube<sup>TM</sup> incluye una extensa documentación con detallados ejemplos para poder desarrollar nuevo código. Esto permite añadir mejoras a través de aplicaciones de terceros, gracias a una Interfaz de Programación de Aplicaciones (*API*) <sup>5</sup> que nos facilita interactuar con otras plataformas. De esta manera se presenta una potente posibilidad de opciones que permiten explotar la información contenida en los metadatos de Youtube<sup>TM</sup> [Benslimane et al., 2008, Wong and Hong, 2007, Fox et al., 2012].

Dado el enorme volumen de vídeos existentes en Youtube<sup>TM</sup> los usuarios pueden sentirse saturados o con mucha información incorrecta o desfasada. Ésto genera una pérdida de tiempo para poder localizar vídeos de calidad, algo particularmente importante en el ámbito de la salud [Fernandez-Luque et al., 2011].

Para poder evaluar las posibles soluciones a este problema, se desarrollaron como parte de este trabajo de Tesis las plataformas Saluteca y Diavideos, cuyo principal objetivo era mostrar la potencialidad del algoritmo HealthTrust [Fernandez-Luque et al., 2012]. Este algoritmo recoge

---

<sup>4</sup>Youtube<sup>TM</sup> statistics [http://www.youtube.com/t/press\\_statistics](http://www.youtube.com/t/press_statistics) available at <http://www.webcitation.org/6BjAbpLHR>

<sup>5</sup>Youtube API <https://developers.google.com/youtube/>

todos los vídeos de salud de distintos Canales y Playlists validados por la comunidad sanitaria, que después se muestran en una plataforma web independiente. Ofrecen vídeos de calidad orientados al paciente; en el caso de Saluteca son vídeos de cualquier temática exclusivamente en español, en el caso de Diavideos orientados a vídeos de Diabetes. Podemos encontrar su funcionamiento en el flujograma de la figura 3.1.

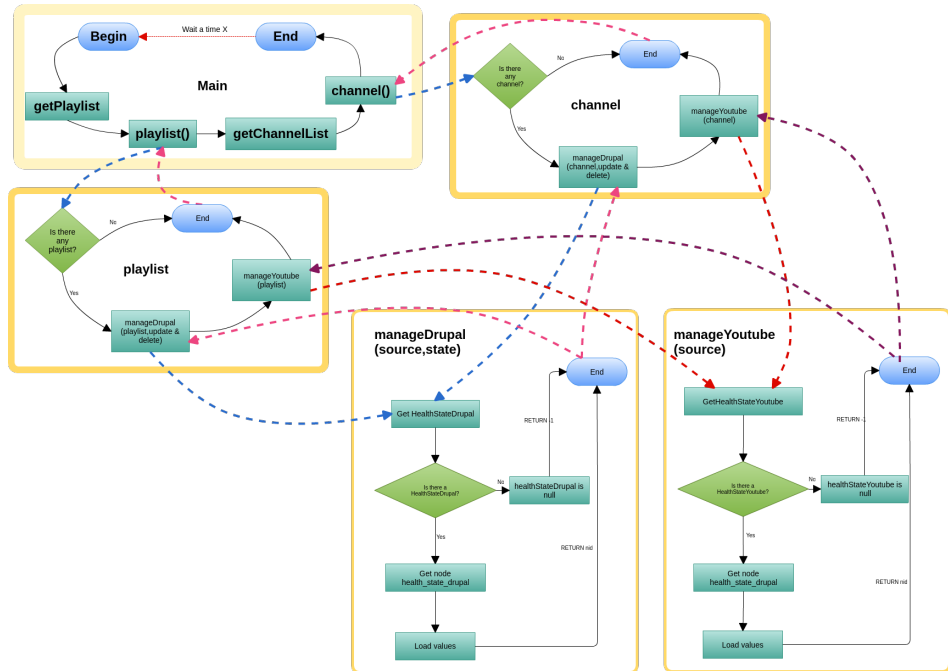


Figura 3.1: Algoritmo de recopilación de vídeos desde Youtube

Las personas pueden conseguir la información que les interesa gracias al filtro de vídeos que ofrecen las plataformas. El funcionamiento conlleva un proceso automatizado que se ejecuta previamente para recoger los vídeos de los canales y playlists. Ambos portales incluyen un desarrollo basado en código abierto que fue desarrollado como parte de esta Tesis durante la estancia en Northern Research Institute (NORUT) <sup>6</sup>.

<sup>6</sup>Northern Research Institute - NORUT <http://norut.no>

Saluteca y Diavideos utilizan la API de Youtube<sup>TM</sup>, así iremos filtrando su contenido de forma continua y periódica. Los elementos seleccionados se incorporan en la plataforma, basada en Drupal CMS<sup>7</sup>.

En el caso de *Saluteca*, el algoritmo gestiona 22 canales válidos:

1. Fundación del corazón.
2. Asociación Española contra el Cáncer.
3. Esclerosis Múltiple.
4. Mundo bebe TV.
5. El blog de Rosa.
6. Ministerio de Salud del Gobierno de Chile.
7. Proyecto salvavidas.
8. Ministerio de la Salud de la Nación.
9. Torre vieja Salud.
10. Fundación Josep Carreras contra leucemia.
11. Escuela de Pacientes.
12. Nos cuidamos.
13. Consejería de Salud de la Junta de Andalucía.
14. Fundiabetes.
15. Hospital San Joan de Déu.
16. Fundación de ayuda contra la drogadicción.
17. Madrid Salud.
18. Salud Andalucía.

---

<sup>7</sup>Drupal web <http://www.drupal.org>

19. Nestle Tv Bebe.
20. Sanidad Cantabria.
21. Ministerio de Sanidad.
22. Servicios Sociales e Igualdad.

En el caso de *Diavideos* son 29 canales y 6 Playlist válidos con un total de unos 1250 vídeos recopilados en continuo incremento:

■ *Canales*

1. Diabetes al día.
2. Federación Diabéticos Españoles.
3. National Diabetes Education Program.
4. Fundación para la Diabetes.
5. International Diabetes Federation.
6. Accu-Chek.
7. World Diabetes Day.
8. American Diabetes Div.
9. DiabeticDanica's channel.
10. Diabetes Recipes.
11. Diabetes Matters.
12. Insulivin.
13. National Diabetes Education Program.
14. Diabetes UK Charity.
15. Australian Diabetes Council.
16. Canadian Diabetes Association.
17. Preop®Diabetes Center.
18. Happy Diabetes.
19. JDRF.
20. DiabetesMine.



21. tudiabetes.
22. Kerry Sparling.
23. Living With MyType2.
24. Joslin Diabetes Center.
25. Diabetes Hands Foundation.
26. ClearlyHealth.com - Patient Ed.
27. Bernard Farrell.
28. dLife.
29. American Diabetes Association.

■ *Playlist*

1. Global Diabetes Symposium.
2. Diabetes by Healthguru.com.
3. Reversing Diabetes by Care World TV.
4. Diabetes by The Doctor's Channel.
5. Type II Diabetes by JNJhealth.
6. Diabetes by JNJhealth.

El funcionamiento del algoritmo reflejado en la figura 3.1 se basa en recoger vídeos de cada canal. Por un lado actualiza los vídeos existentes en la propia plataforma (*ManageDrupal*) y por otro incorpora los nuevos vídeos (*ManageYoutube*). Así tenemos dos estados guardados, uno del último vídeo procesado para ese canal y otro para el último vídeo nuevo procesado para ello:

1. *GetPlayList/GestChannelList*: De la lista de Playlist o Canales obtienen el último explorado o el primero en caso de no haber ninguno.
2. *Playlist()/Channel()*: Iniciamos el análisis de los vídeos para su gestión en nuestra plataforma.
3. *ManageDrupal*: Análisis de vídeos para modificar o eliminar conforme a los datos recabados desde Youtube<sup>TM</sup>

4. *Manage Youtube*: Análisis de los vídeos para incluir los nuevos vídeos a la plataforma.
5. *Wait a time X*: En cada paso guardamos el estado de para el procesado después de un tiempo hasta la siguiente sejecución del algoritmo.

El servidor se configura conforme a las características siguientes: Apache<sup>8</sup>, PHP<sup>9</sup>, Cliente API de Youtube<sup>TM</sup>, Drupal<sup>10</sup> con las adaptaciones de las extensiones que contemple la gestión de contenidos multimedia. El algoritmo escanea automáticamente las fuentes antes descritas e inserta los correspondientes valores (metadatos) del vídeo en el sistema web Drupal. Para ello se programa un proceso en lotes que se ejecuta cada 24 horas. De manera que el sistema llama al algoritmo de la figura 3.1 añadiendo los nuevos vídeos y modificando los existentes. El resultado de Saluteca se ilustra en las figuras 3.2 y 3.3 y en el caso de Diavideos en la figura 3.4.

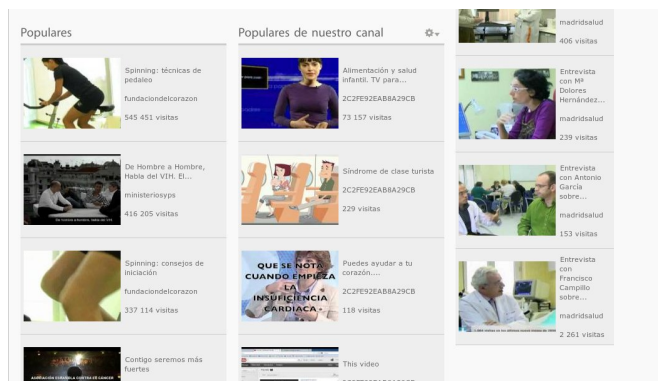


Figura 3.2: Resultado web de Saluteca

Drupal es un sistema gestor de contenido estable que contiene miles de foros con ayuda técnica y de usuario<sup>11</sup>. Además permite múltiples configuraciones, incluida una para la gestión de contenidos digitales multimedia

<sup>8</sup>Apache Software - <http://apache.org>

<sup>9</sup>PHP - <http://php.net>

<sup>10</sup>Drupal - <http://drupal.org>

<sup>11</sup>Drupal forum topics <http://drupal.org/forum>



Figura 3.3: Detalle de un vídeo Saluteca

y otra para el incrustado (embebido) en la propia plataforma de vídeos provenientes de distintos proveedores como Youtube<sup>TM</sup> o Vimeo<sup>TM</sup><sup>12</sup>.

Antes de llevar a cabo la ejecución del algoritmo se tuvo que configurar el entorno Drupal para guardar contenidos multimedia. Éste contiene un espacio con una amplia gama de posibilidades, donde se define la “entidad” como el elemento base de representación de la información. En este caso se crea “*Health Video*” con las propiedades definidas en la tabla 3.1. Además se configuran dos entidades más de *Canales* como puede verse en la tabla 3.2 y *Playlist* apuntado en la tabla 3.3.

Con estas entidades definidas se ha podido llevar a cabo la ejecución del algoritmo que selecciona, actualiza e inserta los vídeos en el portal. El funcionamiento básico del algoritmo 1 realiza toda esta gestión. Para ello se han creado una serie de entidades para poder guardar los estados corres-

<sup>12</sup>Vimeo <http://www.vimeo.com>

|                             |   |
|-----------------------------|---|
| <i>Title</i>                | Contiene el título del vídeo  |
| <i>Video URL</i>            | Url del vídeo en Youtube <sup>TM</sup>  |
| <i>Teaser image</i>         | Imagen a usar como preview  |
| <i>Body</i>                 | Descripción   |
| <i>HealthTrust</i>          | Validación del algoritmo HealthTrust  |
| <i>Enabled</i>              | Control de cambios, en caso que el autor elimine o prohíba su publicación externa se deshabilita  |
| <i>VideoID</i>              | Identificación del vídeo. Cualquier vídeo de Youtube <sup>TM</sup> posee una combinación de 11 caracteres que identifica de forma unívoca al vídeo, está justo despues de la url por ejemplo <code>v=I5Igu1n6bGw</code> <sup>13</sup> |
| <i>upload_date_Youtube</i>  | Fecha de actualización en Youtube <sup>TM</sup>   |
| <i>view_count_Youtube</i>   | Vistas realizadas   |
| <i>Number_votes_Youtube</i> | Votos recibidos   |
| <i>Category_Youtube</i>     | Categoría asignada  |
| <i>Uploader</i>             | Canal o playlist  |
| <i>Tags_Youtube</i>         | Tags ( <i>ya no utilizado</i> )   |
| <i>ChannelID</i>            | Id del canal  |
| <i>HealthVideoType</i>      | Tipo de vídeo   |
| <i>Likes</i>                | Valor de aceptación   |
| <i>Dislikes</i>             | Valor de rechazo  |
| <i>Teaser img1</i>          | Imagen de vista previa  |
| <i>Teaser img2</i>          | Imagen de vista previa  |
| <i>textitTeaser img3</i>    | Imagen de vista previa  |
| <i>Background Image</i>     | Imagen de vista previa de mayor tamaño  |

Tabla 3.1: Entidad HealthVideo

|                          |                             |
|--------------------------|-----------------------------|
| <i>Title</i>             | Título                      |
| <i>Logo</i>              | Logo del Canal              |
| <i>Date joined</i>       | Fecha de creación           |
| <i>lastestActivities</i> | Última actividad            |
| <i>Country</i>           | Ciudad                      |
| <i>city</i>              | País                        |
| <i>about</i>             | Descripción del Canal       |
| <i>ChannelViews</i>      | Vista realizada en el canal |
| <i>UploadViews</i>       | Vistas                      |
| <i>Channel</i>           | Nombre del canal            |
| <i>Channel_url</i>       | Web del canal               |
| <i>Web Site</i>          | Web de proveedor            |
| <i>Body</i>              | Descripción                 |

Tabla 3.2: Entidad HealthChannel

|                             |                        |
|-----------------------------|------------------------|
| <i>Title</i>                | Título                 |
| <i>Logo pl</i>              | Logo del playlist      |
| <i>Date joined pl</i>       | Fecha unión            |
| <i>lastestActivities pl</i> | Última actividad       |
| <i>Country pl</i>           | País                   |
| <i>city pl</i>              | Ciudad                 |
| <i>about pl</i>             | Acerca del Playlist    |
| <i>PlaylistViews pl</i>     | Vistas                 |
| <i>UploadViews pl</i>       | Actualizaciones vistas |
| <i>Playlist</i>             | Nombre del Playlist    |
| <i>Playlist_url</i>         | web del Playlist       |
| <i>Web Site pl</i>          | Web del organizador    |
| <i>Body pl</i>              | Descripción            |

Tabla 3.3: Entidad HealthPlaylist



Figura 3.4: Resultado web de Diavideos

|                            |   |
|----------------------------|---|
| <i>health_state_drupal</i> | Estado puntero en Drupal (para actualización) |
| <i>channel</i>             | Canal asignado al puntero                     |
| <i>element</i>             | Último índice visitado dentro del canal       |
| <i>nidChannelDrupal</i>    | Identificación del canal                      |
| <i>allVideosDrupal</i>     | Contador de todos los vídeos del canal        |

Tabla 3.4: Entidad Estado HealthStateDrupal

pondientes. Uno para los contenidos ya existentes “*Health State Drupal*” descrito en la tabla 3.4 otro para los nuevos contenidos “*Health State Youtube*” expuesto en la tabla 3.5. Estos estados guardan los índices del lugar donde nos quedamos, el canal en el que estamos “channel” o “playlist” y el vídeo último consultado “element” o “start\_index”.

El portal incluye un motor de búsqueda propio de la plataforma Drupal que permite localizar vídeos. El algoritmo 1 procesa 200 videos en cada iteración (100 de actualización y otros 100 de inserción siguiendo el flujograma de la figura 3.1)

La entidad “HealthVideo” representa el vídeo en el portal web. En la figura 3.5 y en 3.6 podemos ver cómo se ha ido incorporando dicha entidad durante las pruebas realizadas en Noviembre del 2013. Se puede observar que el flujo de mayor incorporación está en el pico del 18 de

---

**Algoritmo 1** Recuperación de videos de salud de Youtube™
 

---

- 1: *Pre-selección manual de los canales/Playlist (HealthTrust)*
  - 2: Seleccionar el Canal o Playlist de Youtube™
  - 3: Aplicar el algoritmo HealthTrust
  - 4: **si** es válido **entonces**
  - 5:   Incluir el Canal o Playlist
  - 6: **fin si**
  - 7: *Funcionalidad automática del algoritmo*
  - 8: **repetir**
  - 9:   Recuperar el último estado (para actualizar o insertar)
  - 10: **si** no hay estado **entonces**
  - 11:   Crea un nuevo estado e inicializar
  - 12: **fin si**
  - 13:   Recuperar el canal/playlist e indexarlo
  - 14:   Obtener los siguientes 100 vídeos para actualizar
  - 15:   Obtener los 100 videos para insertar
  - 16:   Guardar las propiedades de los vídeos en el portal
  - 17:   Guardar el nuevo estado
  - 18:   Esperar 30 minutos
  - 19: **hasta que** Infinito
- 

|                             |   |
|-----------------------------|---|
| <i>health_state_youtube</i> | Estado puntero en Youtube (para inserción)        |
| <i>channel</i>              | Canal asignado al puntero                         |
| <i>start_index</i>          | Último índice visitado dentro del canal           |
| <i>nidChannelYoutube</i>    | Identificador del canal                           |
| <i>allVideosYoutube</i>     | Contador de todos los vídeos del canal en Youtube |

Tabla 3.5: Entidad Estado HealthStateDrupal

Noviembre, con más de 800 vídeos procesados. En días posteriores la carga de nuevos vídeos baja para pasar a una situación donde sólo se realizan las actualizaciones correspondientes. Es posible tener un sistema estable en un tiempo aceptable (no más de una semana)

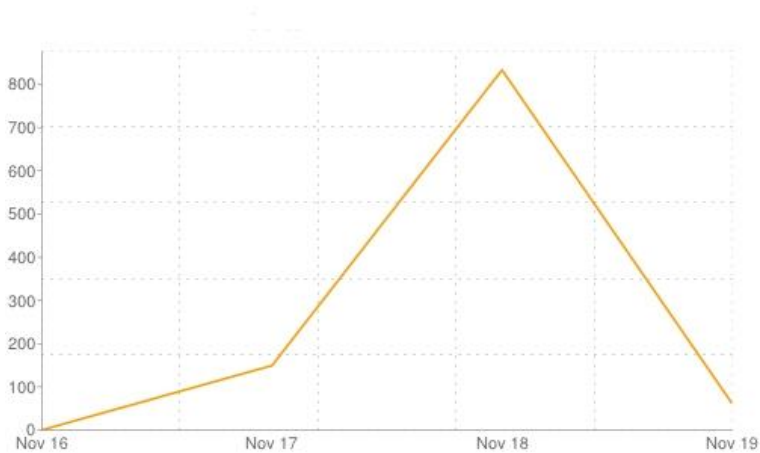


Figura 3.5: Incorporación automática de contenidos en el portal

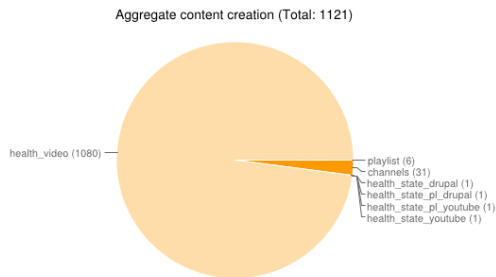


Figura 3.6: Incorporación automática de contenidos en el portal

Las pruebas preliminares han demostrado la utilidad del uso de los sistemas recomendadores dando resultados satisfactorios en las sugerencias de los vídeos [Jäschke et al., 2007]. Después de este piloto de pruebas se



procederá al estudio y adaptación del algoritmo adecuado. Ésto brindará información de salud fiable al paciente gracias a HealthRecSys, método que afina y complementa la información con fuentes válidas de información sanitaria, especialmente orientadas al paciente.

Con esta fase se da solución a los objetivos 1, 2, 3a del capítulo 1.3.

## 3.2. Metodología de Evaluación

Una vez desarrollado el sistema para realizar las pruebas, se debe llevar a cabo la evaluación de los resultados a través de varios ensayos. En ellos tendremos una serie de vídeos con unas recomendaciones que el evaluador (rater) tendrá que valorar si son válidas o no. Para ello se han utilizado dos métodos estadísticos: *Valoración cualitativa de acuerdo entre evaluadores (raters)*, *cálculo de la Precision@k* y *cálculo de ganancia acumulada normalizada nDCG*.

### 3.2.1. Evaluación cualitativa de acuerdo - Kappa Cohen

El coeficiente Kappa Cohen define el nivel de acuerdo entre dos raters [Carletta, 1996], descrito en la fórmula 3.1. El coeficiente  $k$  nos entrega el nivel de acuerdo relativo a una serie de evaluaciones cualitativas, siendo  $Pr(a)$  el acuerdo observado entre los raters y  $Pr(e)$  la probabilidad (estimada) de acuerdo por efecto del azar.

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (3.1)$$

Un valor  $k = 1$  indica un total acuerdo entre los evaluadores. Si  $k > 0,8$  se garantiza una alta concordancia entre raters, es decir, los resultados de ambos raters aceptación/rechazo de enlaces han sido casi coincidentes. De esta forma se garantiza que sólo se utilizan las recomendaciones con suficiente consenso. Nos queda averiguar los resultados de esas recomendaciones, de ello se encarga los puntos 3.2.2 y 3.2.3. La tabla 3.6 muestra los niveles de aceptación.

| <i>Coefficiente Kappa</i> | <i>Fuerza de la concordancia</i> |
|---------------------------|----------------------------------|
| 0,00                      | Pobre                            |
| 0,01 - 0,20               | Leve                             |
| 0,21 - 0,40               | Aceptable                        |
| 0,41 - 0,60               | Moderada                         |
| 0,61 - 0,80               | Considerable                     |
| 0,81 - 1,00               | Casi perfecta                    |

Tabla 3.6: Valoración de coeficiente Kappa

### 3.2.2. Evaluación de las recomendaciones de MedlinePlus - *Precision@k*

Puestos de acuerdo los raters en la aceptación/rechazo de los enlaces necesitamos medir la evaluación de los ítems de la encuesta. La precisión definida en la fórmula 3.2 (también denominada como predicción positiva) establece la fracción de ítems válidos o relevantes de entre todos los recuperados. Una alta precisión reduce el número de falsos positivos (recomendaciones no válidas o relevantes)

$$Precision = \frac{|Recomendacionesvalidas| \cap |Recomendacionesrecuperadas|}{|Recomendacionesrecuperadas|} \quad (3.2)$$

El recobrado (también denominado como recall) se calcula con la fórmula 3.3, en este caso se define como la fracción de ítems recuperados de entre todos los válidos o relevantes. Un recall alto reduce el número de falsos negativos (recomendaciones válidas o relevantes que no son recuperadas)

$$Recobrado = \frac{|Recomendacionesvalidas| \cap |Recomendacionesrecuperadas|}{|Recomendacionesvalidas|} \quad (3.3)$$

La función  $Precision@k(P@k)$  [Büttcher et al., 2016, Kobayashi and Takeda, 2000] se resume en la fórmula 3.4 disponible en [Chapelle and Wu, 2009] donde se considera la precisión de las  $k$  primeras recomendaciones recuperadas,  $P(j)$

es el valor de la precisión de la recomendación  $j$  y  $\Delta r(j)$  es la diferencia entre las recomendaciones  $j - 1$  y  $j$  [Sujatha and Dhavachelvan, 2011].

$$Precision@k = \sum_{j=1}^k P(j) \Delta r(j) \quad (3.4)$$

Cuando el número de recomendaciones válidas es grande el recobrado (recall) no es demasiado útil porque no tiene sentido ofrecerlas todas. En esos casos,  $Precision@k$  nos permite evaluar las  $k$  primeras recomendaciones.

### 3.2.3. Evaluación de las recomendaciones de MedlinePlus - $nDCG$

La ganancia acumulada normalizada  $nDCG$  (normalized Discounted Cumulative Gain) de un sistema recomendador conforme a la fórmula 3.5. Es una medida de la efectividad del sistema que se basa en una escala graduada de relevancia de las entidades recomendadas. Su valor oscila entre 0 y 1, siendo 1 el valor ideal de las entidades [Wang et al., 2013, Valizadegan et al., 2009].

$$nDCG_k = \frac{DCG_k}{iDCG_k} \quad (3.5)$$

Donde  $DCG_k$ , conforme a la fórmula 3.6, evalúan los enlaces relevantes de las primeras posiciones y además enlaces validados que aparecen en posiciones más a finales de la búsqueda, en estos casos se penalizados con el cálculo logarítmico proporcional a la posición.

De esta manera  $DCG_k$  establecen dos suposiciones:

- Los enlaces válidos altamente relevantes son más útiles cuando aparecen antes.
- Los enlaces válidos que aparecen más tardíamente en la búsqueda son más útiles que los enlaces rechazados.

$$DCG_k = \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}, rel_i \in 0, 1 \quad (3.6)$$

Por otro lado  $iDCG_k$ , referenciado en la fórmula 3.7 representa el vector ideal, en nuestro caso consideramos este valor aquél que todos los  $k$  enlaces recomendadores son válidos.

$$iDCG_k = \sum_{i=1}^k 1 \quad (3.7)$$

### 3.3. HealthRecSys

HealthRecSys es el algoritmo que filtra información sobre un vídeo y usa dicha información para recomendar contenidos complementarios de fuentes de salud fiables como: MedlinePlus<sup>14</sup>, SNOMED-CT<sup>15</sup> entre otras. Una de las principales aportaciones de esta tesis es precisamente el análisis, diseño e implementación de este algoritmo. A lo largo del proceso se introdujeron diversas mejoras que permitieron complementar la información que ayuda a la búsqueda de recomendaciones.

Se pueden identificar cuatro etapas principales que describiremos a continuación:

- *Etapla previa - Etiquetas*: Esta fue la línea primera que planteamos en la elaboración del algoritmo, pero que tuvo que dejarse apartada por limitaciones de la API de Youtube<sup>TM</sup>.
- *Etapla primera - Términos para recomendar*: En esta etapa nos centramos en el estudio de la palabra o conjunto de palabras (términos) para la obtención de resultados recomendadores.
- *Etapla segunda - Ontologías*: Se incorporó a la etapa anterior (términos) la búsqueda a través de ontologías sanitarias.
- *Etapla tercera - Procesado de Lenguaje Natural*: Finalmente se añadieron las funcionalidades semánticas del procesado de lenguaje natural.

A continuación detallaremos más concretamente cada una de estas etapas.

---

<sup>14</sup>MedlinePlus <https://medlineplus.gov/>

<sup>15</sup>SNOMED-CT <http://www.ihtsdo.org/snomed-ct>

### 3.3.1. Etapa previa - Etiquetas

Siguiendo la base del estudio de Diavideos, y observando el gran volumen de vídeos recopilados (más de 1000), se vio la necesidad de mejorar los resultados para poder ofrecer información personalizada al usuario. El contenido de los vídeos en sí podría ser incompleto, con información que se da por conocida. La persona podría llegar a confundir o no entender el mensaje del vídeo ya que se contextualiza en un conocimiento previo que el usuario no necesariamente posee. Por poner un ejemplo, podría ocurrir que hubiera pacientes que no entendieran qué diferencia hay entre la Diabetes Tipo I o Tipo II.

Ante esa carencia de información, se vio la necesidad de complementar con contenido recomendador que dé refuerzo a la información del vídeo en sí. Para ello, y teniendo en cuenta que nuestro objetivo es mejorar la información al paciente, el trabajo se enfocó en el uso de la fuente de referencia MedlinePlus <sup>16</sup>.

El objetivo en este punto era ofrecer recomendaciones de MedlinePlus relativos al contenido del vídeo que el paciente estuviera viendo. De esta manera se podrían reforzar los conceptos que en el vídeo se exponían y ayudar así al paciente a alcanzar una idea más cercana a lo que se pretende informar.

En el capítulo 2 se han expuesto los distintos tipos de sistemas recomendadores existentes: desde los clásicos a los más avanzados. Antes de poder decidir el tipo de sistema recomendador se tuvo que analizar el contexto y las limitaciones para asociar el tipo recomendador más adecuado.

Entre las limitaciones de nuestro sistema HealthRecSys están:

1. La plataforma no reconoce perfiles de usuarios como los que mencionamos en la secciones 2.1.1 y 2.5.1. El recorrido por los vídeos se hace de forma anónima, por lo que no podemos guardar información asociada al usuario que nos ayudaría a poder dar una recomendación más certera.
2. Los metadatos incorporados en la información del vídeo pueden ser vagos e inexactos, incluso a veces confusos. Como veremos, uno de

---

<sup>16</sup>MedlinePlus web <https://medlineplus.gov/>

nuestros principales objetivos será depurar esta información para mejorar las recomendaciones.

3. Los contenidos encontrados se enfocan a recomendaciones de Medline-Plus, referencias de página descriptivas de un concepto que debemos extraer del vídeo. Esto nos lleva a dar recomendaciones con términos más fáciles de extraer, pero perdiendo otros (huérfanos) que pueden también resultar interesantes de mostrar.

Teniendo en cuenta estas circunstancias se tuvo que pensar en sistemas recomendadores basados en el contexto, descritos en la sección 2.2.3. La pérdida de información relativa al perfil del usuario es un punto débil para encontrar las recomendaciones más apropiadas. En un principio nuestro trabajo se enfocó en analizar los metadatos de Youtube<sup>TM</sup> y poder reconocer así aquellos elementos que ayudaran a tener el mejor resultado. Para ello se observaron las etiquetas asociadas al vídeo como una de las primeras opciones. Éstas permiten dar una serie de valores que acercará a conceptos asociados a enlaces MedlinePlus [Konstantinidis et al., 2013].

Los resultados preliminares del trabajo fueron esperanzadores, hasta que en 2012 Youtube<sup>TM</sup> decidió retirar el uso de etiquetas (tags) de la API de Youtube<sup>TM</sup><sup>17</sup>. Esto supuso un revés en la línea de investigación ya que era la base del desarrollo de nuestro sistema recomendador. Se tuvo que hacer un cambio de rumbo abordando otros metadatos que sí fueran accesibles por la API.

### **3.3.2. Etapa primera - Términos para recomendar**

Esta etapa primera se centra en la búsqueda de términos entre los restantes metadatos. Para ello se definió el concepto de término como una palabra o conjunto de palabras que pueden ser representativos en el ámbito de la salud. Del volumen de información obtenida en los títulos y descripción se observa que algunos conjuntos de palabras forman términos pertenecientes a definiciones del área sanitaria. El objetivo en este punto es poder

---

<sup>17</sup>Eliminación de tags desde la API de Youtube - <http://www.genbeta.com/web/youtube-oculta-las-etiquetas-de-todos-sus-videos-para-evitar-las-malas-practicas-de-los-usuarios>

obtener términos de salud de los metadatos de Youtube<sup>TM</sup> y poder filtrarlo a fuentes reconocidas como MedlinePlus.

El primer artículo publicado [Rivero-Rodriguez et al., 2013] se centró en poder extraer los términos de salud de los metadatos del título y descripción del vídeo. Éste fue el primer acercamiento a una recomendación en función de términos, para lo que se aplicaron 3 métodos A,B,C y D (ver figura 3.7)



Figura 3.7: Algoritmo de recomendación propuesto

El método A lo que hace es aplicar un filtro básico de palabras simples sin valor (*stopwords*) como preposiciones, adverbios, etc que son fáciles de discriminar. Las restantes son consideradas como un término en salud que se inyecta en la API de MedLinePlus para ver su resultado, dando lugar a

lo que representa como *Out A*.

El método B añade más filtros al anterior método usando SNOMED-CT, que es una terminología clínica integral y multilingüe de mayor precisión. Incluye un extenso volumen de términos que representa información clínica, y dispone de su propia API. Su resultado se puede inyectar sobre la API de MedlinePlus. En este caso sólo se aplica una palabra como término dando su consecuente resultado como *Out B*.

El método C es igual que el anterior salvo que se considera un conjunto de palabras como un término para inyectarlo a la API de MedlinePlus dando lugar a la salida *Out C*. Por último el método D es la unión de la ejecución de B y C.

En los resultados se aplicó la evaluación descrita en la sección 3.2 donde se obtiene la precisión del punto 3.2.2. Se concluyó que las buenas recomendaciones en el método A y C/D eran similares, aunque el volumen de resultados era mayor en A (con un gran porcentaje de recomendaciones incorrectas) como indica la Figura 3.8. Podemos comparar los resultados entre *method A* y *method D* con la valoración de resultados válidos, que muestran su escaso resultado (por debajo de 25 enlaces recomendados)

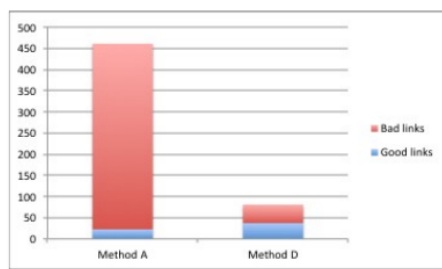


Figura 3.8: Resultados obtenidos de la etapa 3.3.2

El estudio garantiza resultados prometedores con el uso de la API de SNOMED-CT para dar recomendaciones de MedlinePlus. Aunque una limitación es que usa exclusivamente el título del vídeo, que en muchos casos no representa la información propia del vídeo. Por otro lado el escaso número de enlaces válidos de MedlinePlus obliga a mejorar la técnica de recomendación para incrementar el valor de los resultados de recomendación asociado al vídeo.



Se ha logrado llegar a un recomendador prometedor pero que no entrega suficientes resultados como para poder considerarlo interesante. Además los métodos de evaluación llevados en este estudio no fueron los más adecuados: entre otras cosas no se realizó un estudio de concordancia del punto 3.2.1 que permitiera una validación apropiada.

Esta fase dio solución a los objetivos planteados en la disertación correspondientes al punto 3b y 4 de la sección 1.3. El paso siguiente viene a corregir estos puntos débiles y mejorar la fase de evaluación.

### 3.3.3. Etapa segunda - Ontologías

Haciendo referencia a la sección 2.3.3, quedan pendientes las siguientes debilidades:

1. Añadir más recomendaciones.
2. Ampliar el rango de metadatos del contenido del vídeo.
3. Incorporar una evaluación más adecuada.

Además del *título* se incorporó la *descripción* del vídeo en el procedimiento de recomendación, con el objetivo de corregir la debilidad 2. La estrategia se cambió usando los términos ontológicos existentes en Salud. Uno de los bancos de datos más grandes de estos términos lo tenemos en Bioontology<sup>18</sup> que dispone de más de 600 ontologías clínicas entre ellas SNOMED-CT y MedlinePlus. Esto facilita añadir más personalización y explotar mejor el campo de actuación en la búsqueda de recomendaciones [Fernandez-Luque et al., 2012].

Durante esta etapa de realización de la tesis tuvo lugar la concesión de la beca Yggdrassill del Norwegian Research Council<sup>19</sup> con la aprobación del proyecto *Diabetes HealthRecSys*. El proyecto buscaba crear un sistema recomendador en salud para diabetes. Se mantuvo el planteamiento de un

---

<sup>18</sup>Bioontology portal <http://www.bioontology.org/>

<sup>19</sup>El programa Yggdrassill otorga becas para las instituciones de investigación noruegas (universidades, institutos de investigación), en este caso se aplicó para la Universidad de Tromsø (<https://uit.no>) en colaboración con Northern Research Institute - NORUT (<http://norut.no>)-<http://www.forskningsradet.no/en/Funding/ISMOBIL/1233557743178>

caso de estudio y se realizó una demostración aplicada del mismo sobre Diavideos.

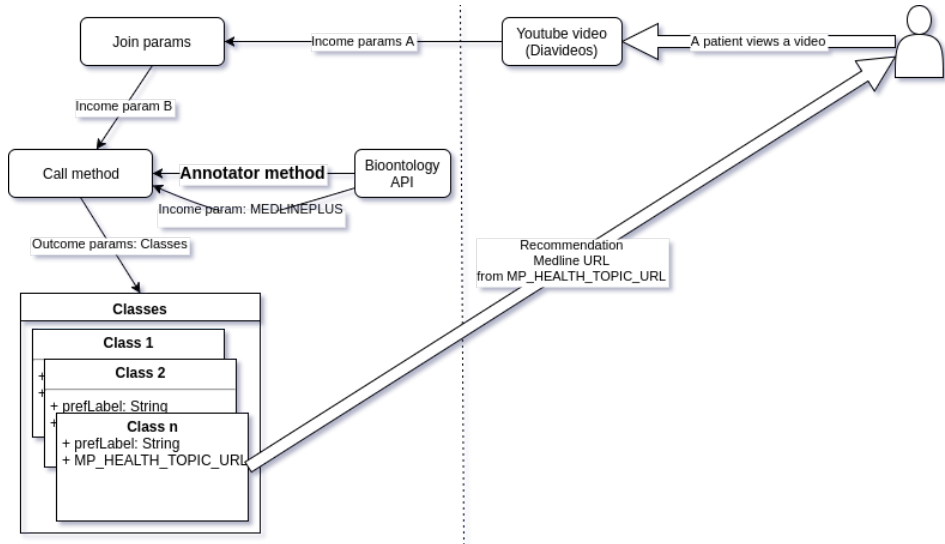


Figura 3.9: Algoritmo de recomendación HealthRecSys de la sección 3.3.3

Los hitos en el proyecto marcados son:

- H1. Analizar los contenidos y metadatos.
- H2. Actualizar y mejorar el framework de Diavideos.
- H3. Estudiar los resultados y compararlos con trabajos previos.

El sistema generado a través de esta beca mejora al anterior, dichos resultados fueron publicados en [Fernandez-Luque L., 2014]. En este caso el método desarrollado fue aplicado en dos ámbitos de la salud: Diabetes e Hipertensión/HTA. Se escogieron 23 vídeos (con un total de 114 recomendaciones) repartidos de la siguiente manera con la condición que mostraran más de 3 recomendaciones:

- 6 de Diabetes recogidos de Diavideos.

- 17 de Presión Arterial de los vídeos más populares de Youtube<sup>TM</sup> aplicando el algoritmo HealthTrust.

El método se ajusta conforme a la figura 3.9. En principio tiene el mismo inicio que en la etapa 3.3.2: quitar las palabras no significativas (*stop words*) y unir las restantes adecuadamente para conseguir términos. Éstos serán la entrada *income param B* para la llamada a la API de Bioontology que contiene el método *Annotator*, explicitando que use la ontología *MedlinePlus* como se muestra en los algoritmos 2 y 3. En ella se incluye la cadena de la API incorporando la *APIKEY*<sup>20</sup>.

---

**Algoritmo 2** Algoritmo recomendador

---

**Entrada:** Lista de vídeos de Youtube<sup>TM</sup>

**Salida:** Array de recomendaciones MedlinePlus (url)

url  $\leftarrow$  Array nuevo

**para** (Obtener cada video) en *text* **hacer**

urltext  $\leftarrow$  getBioOntologyData(text); //Añadir al array el valor de la funcion entrada text

**fin para**

devolver url

---

La salida ofrece un grupo de clases, cada una vinculada a una serie de términos *prefLabel*. En el caso de la ontología MedlinePlus entrega un valor en la clase que hace referencia a un enlace web de MedlinePlus relacionado con el término *MP\_HEALTH\_TOPIC\_URL*. Es precisamente ese enlace el que se suministra como recomendación. Así se obtienen enlaces recomendadores precisos relacionados con el vídeo. En este caso, adelantando algunos datos de la evaluación con raters o evaluadores profesionales sanitarios que se describirá en detalle más adelante, se obtuvieron los resultados en la tabla 3.7 aplicando coeficiente de Kappa Cohen para encontrar la concordancia entre los evaluadores, tal y como expresamos en la sección 3.2.1.

Con este nuevo método hemos ajustado los resultados más objetivamente corrigiendo la debilidad del punto 3 de esta misma sección: por un lado

---

<sup>20</sup>API Bioontology - [http://data.bioontology.org/annotator?text=\\$text &maxlevel=5&apikey=YourAPIKEY&maxlevel=5 &ontologies=MEDLINEPLUS](http://data.bioontology.org/annotator?text=$text &maxlevel=5&apikey=YourAPIKEY&maxlevel=5 &ontologies=MEDLINEPLUS)

---

**Algoritmo 3** Función `getBioOntologyData`

---

**Entrada:** Texto proveniente del título o descripción**Salida:** Un array de enlaces recomendados MedlinePLus (urls)in  $\leftarrow$  "Cadena de API Bioontology"json  $\leftarrow$  `Obten_contenido(in)`data  $\leftarrow$  `Decodifica_json(json, true)`urls  $\leftarrow$  Array limpio**para** (Cada texto) en *value* **hacer**    urlClass  $\leftarrow$  value[.annotatedClass][links][self]; Obtener el valor anotador    urls []  $\leftarrow$  `getUrlClass(urlClass)`; Añadir el valor al array**fin para**devolver urls

---

|                         |   |
|-------------------------|---|
| <i>Diabetes</i>         | Precision@3 de 0.67<br>Kappa Cohen 0.65 |
| <i>Presión Arterial</i> | Precision@3 de 0.91<br>Kappa Cohen 0.65 |

Tabla 3.7: Resultados análisis Algoritmo HealthTrust de la sección 3.3.3

con la selección de raters pertenecientes al sector salud con conocimientos de las nuevas tecnologías, por otro aplicando el coeficiente de concordancia entre dos raters (Kappa Cohen). En el estudio se mostró valores moderados de acuerdo con precisiones aceptables. El principal problema que ofrecía este método está en el reducido volumen de enlaces recomendadores en muchos de los vídeos, con sólo 1 o 2 recomendaciones.

En este estudio se vio una gran oportunidad de cómo explotar la ontología sanitaria para poder dar recomendaciones de calidad. Se reforzó la búsqueda de la máxima concordancia entre dos raters ( $p > 0,8$ : máximo acuerdo) en la sección 3.2.1. No obstante no se pudieron llegar a esos acuerdos con valores  $p = 0,65$  en Diabetes e Hipertensión/HTA. Además se intentó poder incrementar el número de valores recomendadores conforme a la debilidad 1 de esta misma sección, sólo se escogieron los vídeos que mostraban más de 3 recomendaciones de acuerdo con la necesidad del punto 3.2.2, los restantes se tuvieron que descartar.

Esta fase dio solución a los objetivos planteados en la disertación correspondientes al punto 3b, 4 y 6 de la sección 1.3. La idea de continuar explorando el potencial de uso de la semántica y las ontologías es algo que a día de hoy seguimos manteniendo. En el siguiente estudio, se busca afinar la obtención de más términos que encajen con los de la ontologías.

### 3.3.4. Etapa tercera - Procesado de Lenguaje Natural- NLP

Se necesita una respuesta a la cuestión: *¿Cómo incrementar el número de recomendaciones sin perder la calidad?*. Se estudió la idea de aflorar términos ocultos (huérfanos) de MedlinePlus que puedan incrementar el valor de la recomendaciones sin el menoscabo de la calidad del sistema anterior. Tras meses de dudas y continuas pruebas con otras estrategias de recomendación se decidió apostar por el uso del procesado de lenguaje natural (NLP) detallado en la sección 2.5.3.

Para poder incluir NLP en la mejora del algoritmo debemos encontrar una implementación que pueda permitir integrarlo dentro del flujo del procedimiento recomendador, para ello se encontró el framework *cTAKES*<sup>21</sup> que es un proyecto de la licencia Apache<sup>22</sup> dentro de Apache Software

---

<sup>21</sup>Web cTakes <http://ctakes.apache.org/>

<sup>22</sup>Apache licence - <http://www.apache.org/licenses/LICENSE-2.0>

Foundation <sup>23</sup>. Es un sistema de procesado de lenguaje natural relativa a información médica y clínica [Savova et al., 2010].

Con cTAKES podremos extraer información clínica médica, desarrollada inicialmente por Mayo Clinic. Tiene una fuerte y amplia documentación técnica como guía de uso. Su potencialidad ha hecho que haya sido incorporada a sistemas de diversas instituciones internacionales de prestigio. Entre otras entidades podremos encontrar: Hospital Infantil de Boston, Mayo Clinic, Instituto Tecnológico de Massachusetts, Universidad de Pittsburg, Universidad de California.

Dada su solidez y dinamismo, con cambios diarios, se estudió la inclusión en HealthRecSys. cTAKES representa la información en un árbol de nodos (compatible con el formato XML). Para nuestro sistema se seleccionó la rama puramente semántica, como ejemplo podemos ver la figura 3.10<sup>24</sup> donde se detallan los metadatos que contemplan un nodo. Las propiedades que lo albergan son predefinidas por el propio framework [?, Liu et al., 2013].

En dicha figura tenemos las correspondencias de cTakes con HL7. Un mapeo de los atributos del modelo de la arquitectura clínica documental (clinical document architecture model/ CDA model) de HL7 que cruza con los metadatos con el análisis de estructura común (Common Analysis Structure Model - CAS Model) de cTakes.

El concepto *cas:Concept:conceptType:problem* indica un problema o una prueba, lo mapeamos en *cda:ClinicalStatement:classCode:observation*. En el mismo sentido mapeamos *cas:Concept:conceptType:test* con *cda:ClinicalStatement:classCode:procedure*.

cTAKES permite extraer la parte semántica de documentos a partir de las representaciones de las palabras usando técnicas de inteligencia artificial. Este proceso brinda una comprensión del significado en textos complejos, dando resultados en áreas como el análisis de sentimientos [Daniulaityte et al., 2016], la traducción automática [Turner et al., 2015], o en la detección de entidades [Florin et al., 2003].

El método llevado a cabo en este caso consiste en generar enlaces webs de MelinePlus usando los términos ontológicos de Bioontology a partir de

<sup>23</sup>Apache Software Foundation <http://www.apache.org/>

<sup>24</sup>Fuente cTakes: <http://jamia.oxfordjournals.org/content/early/2014/11/07/amia.jnl-2014-002991>

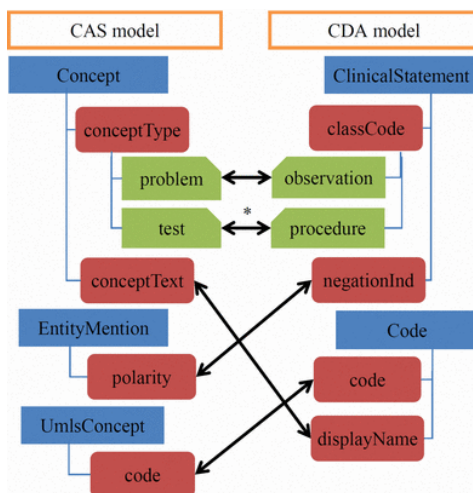


Figura 3.10: Ejemplo de modelos CAS - fuente: cTAKES

un documento -descripción, título y subtítulos de un vídeo de Youtube<sup>TM</sup>-. Se ha estructurado en dos niveles independientes, cada uno con algoritmos diferentes que se adecúan a la búsqueda del término (ver figura 3.11). Esa independencia permite poder añadir más niveles que puedan afinar la búsqueda así como reducirla conforme a la petición del contexto recomendador.

## Procedimiento

**Primera fase: Diseño del algoritmo** En la mencionada figura 3.11 podemos encontrarnos con el módulo *Extract terms* encargado de extraer y filtrar términos desde Youtube<sup>TM</sup>. Los dos niveles descritos son:

- *Nivel 1 - Palabras como términos:* El título, descripción y subtítulo del vídeo se recogen como posibles términos para inyectarlos a la API de Bioontology. Dado que tiene más de 600 ontologías se decidió escoger la ontología de MedlinePlus.
- *Nivel 2 - Procesado de lenguaje natural:* Se utiliza NLP en título, descripción y subtítulos como en el Nivel 1 pero en este caso se aplica

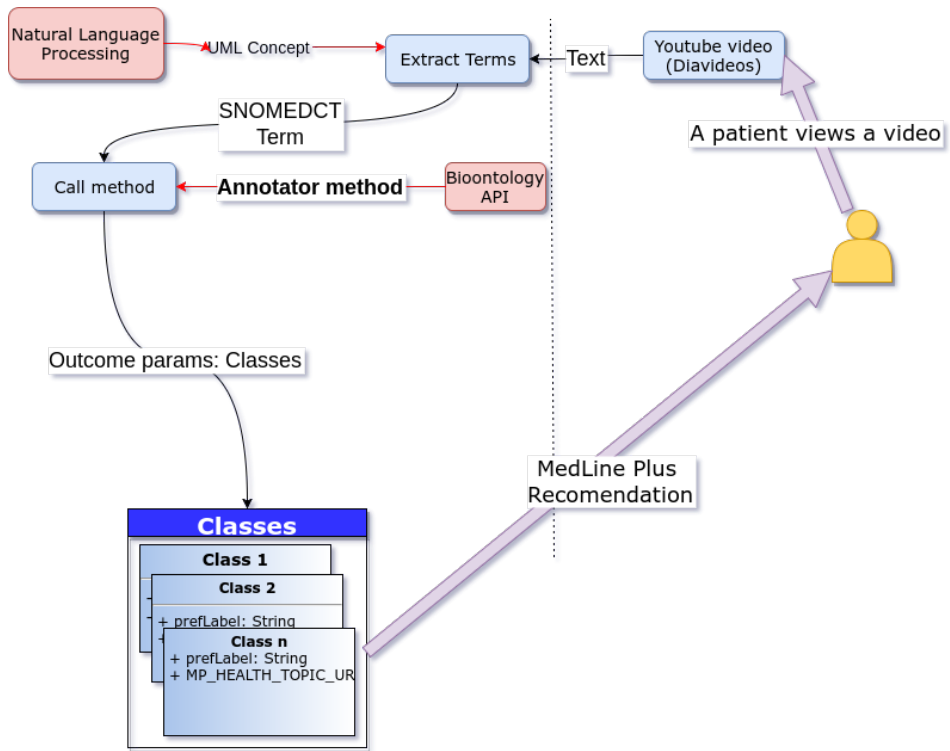


Figura 3.11: Nuevo modelo de HealthRecSys de la sección 3.3.4

un filtro previo con el framework de cTAKES -como se ha indicado previamente una implementación NLP que extrae términos de SNOMED-CT desde el texto-. El término generado se incluye como entrada a la API de Bioontology asignando la ontología MedlinePlus.

Se ha llevado a cabo el análisis de texto utilizando el Sistema Unificado de Lenguaje Médico (UMLS 2) [Bodenreider, 2004] con anotaciones SNOMED-CT para que coincida con el framework cTAKES. La biblioteca UMLS requiere que nos registremos, el apéndice A muestra la configuración completa.

Tal y como se menciona, el texto se incorpora en dicho framework para dar lugar a un archivo XML con el análisis morfológico, sintáctico



y semántico. A partir de este archivo se filtran las etiquetas *UmlConcept* que contienen los términos recogidos desde la ontología SNOMEDCT. Por ejemplo, en la fuente XML 3.1 se muestra extractos de los términos *Blood*, *Entired Cell* y *Cells* recogidos de la entidad *org.apache.ctakes.typesystem.type.refsem.Umlsconcept* definido en el documento XML.

Una vez que se extraen los términos SNOMED-CT, se cruzan con la API de Bioontology para encontrar términos homónimos en la ontología MedlinePlus. Esos parámetros de salida permiten obtener un enlace web con la propiedad *MP\_HEALTH\_TOPIC\_URL*, lo que ofrece información fiable de la web de MedlinePlus. Siguiendo el ejemplo anterior, los términos *Blood*<sup>25</sup>, *Entired Cells* y *Cells*<sup>26</sup> corresponden con sus webs correspondientes de MedlinePlus.

El número de términos de vocabulario SNOMED-CT es superior a los de MedlinePlus. Como se preveía que muchos resultados no tendrían términos coincidentes, en esos casos, los términos SNOMED-CT serán ignorados pasando a analizar el siguiente término. Aunque no se descarta que pueda cruzarse con otras ontologías, en esta etapa se ha centrado en los resultados de MedlinePlus que están orientados a dar información al paciente.

Fuente XML 3.1: Sección fichero configuración cTAKES

```
<org.apache.ctakes.typesystem.type.refsem.UmlsConcept _id="24523"
  codingScheme="SNOMEDCT" code="87612001" oid="87612001#SNOMEDCT"
  score="0.0" disambiguated="false" cui="C0005767" tui="T024"
  preferredText="Blood" />
<uima.cas.FSArray _id="24533" size="1">
  <i>24523</i>
</uima.cas.FSArray>
<org.apache.ctakes.typesystem.type.refsem.UmlsConcept _id="24488"
  codingScheme="SNOMEDCT" code="362837007" oid="362837007#SNOMEDCT"
  score="0.0" disambiguated="false" cui="C1269647" tui="T025"
  preferredText="Entired Cell" />
```

<sup>25</sup>Blood - <https://www.nlm.nih.gov/medlineplus/blood.html>

<sup>26</sup>Cell - <https://www.nlm.nih.gov/medlineplus/stemcells.html>

```
<org.apache.ctakes.typesystem.type.refsem.UmlsConcept _id="24478"
  codingScheme="SNOMEDCT" code="4421005" oid="4421005#SNOMEDCT"
  score="0.0" disambiguated="false" cui="C0007634" tui="T025"
  preferredText="Cell" />
```



Title: Diabetes Foot Exam  
Description  
One in four people with diabetes will develop foot problems that require treatment. You can develop different types of foot problems, but all can lead to serious complications if left untreated. Richard A. Jackson, M. D., of the Joslin Clinic at the Joslin Diabetes Center, explains regular foot exams are important to people with diabetes. Learn more about Joslin's Foot Center at [http://www.joslin.org/care/foot\\_care...](http://www.joslin.org/care/foot_care...)

|                      | Accept                | Reject                |
|----------------------|-----------------------|-----------------------|
| Diabetes Mellitus    | <input type="radio"/> | <input type="radio"/> |
| Infections           | <input type="radio"/> | <input type="radio"/> |
| Smoking              | <input type="radio"/> | <input type="radio"/> |
| Blood                | <input type="radio"/> | <input type="radio"/> |
| Cholesterol          | <input type="radio"/> | <input type="radio"/> |
| Communicable Disease | <input type="radio"/> | <input type="radio"/> |
| Calcium              | <input type="radio"/> | <input type="radio"/> |

Figura 3.12: Demo de la encuesta

### 3.4. Evaluación de HealthRecSys

Para mejorar la evaluación en la sección 3.2 con respecto a la anterior etapa se han incrementado tanto los vídeos como los raters. Se escogieron 53 vídeos y se invitaron a 26 raters para cumplimentar una encuesta de los mencionados vídeos (ver ejemplo en la figura 3.12).

Los profesionales sanitarios colaboradores fueron invitados por su familiaridad con la salud y el conocimiento de los medios sociales. Se pidió que determinaran si los enlaces recomendados para un vídeo determinado eran relevantes para el tema que trata. Para seleccionar la clasificación de la relevancia de los vídeos y los enlaces recomendados usamos la métrica Kappa Cohen para determinar el nivel de acuerdo entre dos evaluadores. Se busca

un coeficiente Kappa cuya valoración sea buena o muy buena ( $k > 0,6$  a  $k < 0,91$ ) para garantizar el máximo acuerdo entre pares conforme a la evaluación de la sección 3.2.1. El reparto se realizó de la siguiente forma:

- 10 de los vídeos del artículo [Rivero-Rodriguez et al., 2013]. 2 raters terminaron la encuesta. Llamamos a este caso CBMS por el acrónimo del congreso donde se presentó.
- 43 videos escogidos de la siguiente forma:
  - 22 correspondientes a Diabetes (incluyendo los vídeos del artículo [Fernandez-Luque L., 2014] recopilados de Diavideos). 3 raters terminaron la encuesta.
  - 21 correspondientes a Hipertensión/HTA aplicando el algoritmo del artículo [Fernandez-Luque et al., 2012] recopilados del propio Youtube<sup>TM</sup>. 7 raters terminaron la encuesta.

Usando nuestro algoritmo generamos 251 recomendaciones de enlaces de MedlinePlus. Los resultados se muestran en las tablas 3.8 y 3.9. Todos muestran un grado de concordancia referido en la sección 3.2.1 entre el 60 % y 95 % en las tres y cuatro primeras recomendaciones ( $Prec@3$  y  $Prec@4$ ) para las muestras estudiadas. En nuestro caso un coeficiente Kappa Cohen de 0,626 para CBMS ( $z = 4,35$ ,  $p\text{-value} = 1,33 \times 10^{-5}$ ), 0,582 para Diabetes ( $z = 6,47$ ,  $p\text{-value} = 9,9 \times 10^{-11}$ ) y 0,717 para Hipertensión ( $z = 7,72$ ,  $p\text{-value} = 1,31 \times 10^{-14}$ ).

El cálculo de la  $Precision@k$  y  $nDCG$  se lleva a cabo a partir de los videos filtrados previamente:

- Dado un video  $i$ , la  $Precision@k$  y  $nDCGk$ , de las primeras  $k$  recomendaciones de enlaces, se obtiene de la siguiente manera -para cada enlace  $j$  ( $1 \leq j \leq k$ ) en  $recom_j$ ,
  - a Si ambos evaluadores aprueban el enlace  $j$ , se toma por aceptado (valor 1).
  - b Si ambos evaluadores no aprueban el enlace  $j$ , se toma por rechazado (valor 0).
  - c En caso de discrepancias, que un evaluador acepte el enlace  $j$  y el otro lo rechace.

- 1 Caso robusto: (AND) - se considera rechazado.
- 2 Caso moderado: (OR) - se considera aceptado.

El cálculo del  $Precision@k$  para el vídeo es la relación entre los valores obtenidos previamente, dando lugar a dos resultados en los casos de desacuerdo. En la tabla 3.8 el cálculo de las  $Precision@k$  fueron realizadas manualmente con la herramienta Excel. En la tabla 3.9 queda reflejado el cálculo  $DCG_k$  para el cual se usó el programa R conforme al algoritmo 4. Se toma como  $iDCG_k$  al  $DCG$  del vector de enlaces ideales, es decir aquellos que todas las recomendaciones son aceptadas (valor 1). Las figuras 3.15, 3.14 y 3.13 muestran gráficamente los valores reflejados en las tablas mencionadas.

| Mean<br><b>Precision@k</b> | Robust case |            |            | Moderate case |            |            |
|----------------------------|-------------|------------|------------|---------------|------------|------------|
|                            | <b>p@3</b>  | <b>p@4</b> | <b>p@5</b> | <b>p@3</b>    | <b>p@4</b> | <b>p@5</b> |
| <b>CBMS</b>                | 0,77        | 0,65       | 0,5        | 0,87          | 0,8        | 0,7        |
| <b>Diabetes</b>            | 0,71        | 0,71       | 0,68       | 0,89          | 0,85       | 0,81       |
| <b>Hypertension/HTN</b>    | 0,48        | 0,45       | 0,39       | 0,62          | 0,57       | 0,53       |

Tabla 3.8: Cálculo de las medias de  $Precision@k$  de enlaces recomendados

| Mean nDCG              | Robust case |            |            | Moderate case |            |            |
|------------------------|-------------|------------|------------|---------------|------------|------------|
|                        | <b>p@3</b>  | <b>p@4</b> | <b>p@5</b> | <b>p@3</b>    | <b>p@4</b> | <b>p@5</b> |
| <b>CBMS</b>            | 0,78        | 0,7        | 0,5        | 0,88          | 0,83       | 0,75       |
| <b>Diabetes</b>        | 0,73        | 0,74       | 0,72       | 0,90          | 0,87       | 0,85       |
| <b>Hypertesion/HTN</b> | 0,51        | 0,49       | 0,46       | 0,65          | 0,61       | 0,58       |

Tabla 3.9: Cálculo de las medias de nDCG de enlaces recomendados

Este estudio demostró que un algoritmo puede proporcionar recomendaciones de contenido relevantes en la recomendación de tercer o cuarto nivel para vídeos relacionados con CBMS, Diabetes e Hipertensión. La precisión cae significativamente después de la quinta recomendación, lo que plantea la cuestión de cuántas recomendaciones son suficientes para un paciente.

Las mejoras futuras de los sistemas de recomendación incorporarán más análisis semántico y tal vez puedan determinar el estado de ánimo del paciente para hacer mejores recomendaciones.

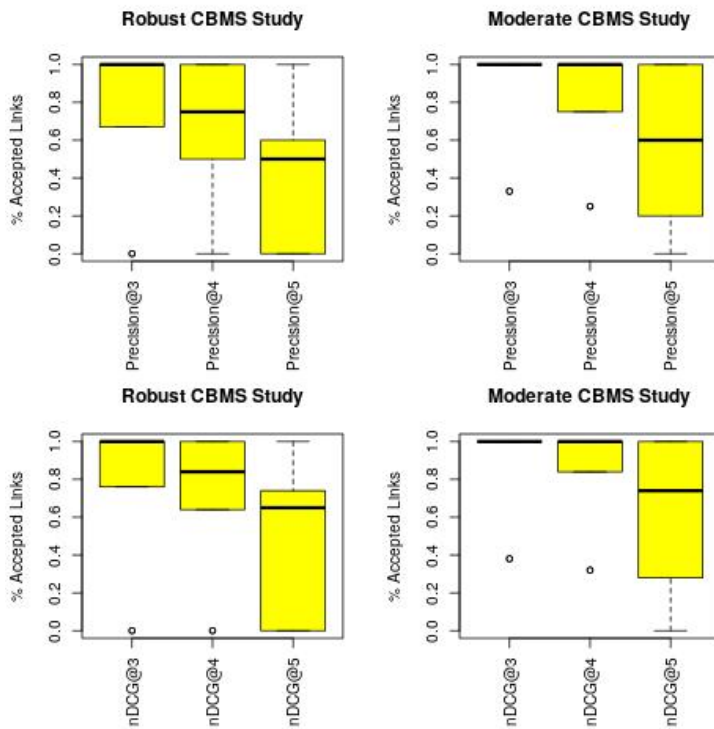


Figura 3.13: Resultados en CBMS

Por otro lado, el sistema extrae información implícita -no presente en el contenido de los metadatos-. Esta técnica incluye términos similares (vecinos) que pueden ser susceptibles de recomendación. En este caso se escogieron los vecinos de primer grado (justo el más cercano). La forma de selección de los vecinos lo encontramos con el fichero de resultados que nos entrega sinónimos al cual podemos cruzar términos. El primer nivel pararía en la primera fase de sinónimos sin abordar los subsiguientes términos. Por poner un ejemplo, el vídeo *Juvenile diabetes Research Foundation - Cure Video - Dalas*<sup>27</sup> (ver figura 3.16) muestra el impacto de la diabetes en jóve-

<sup>27</sup>Juvenile Diabetes Research Foundation - <https://www.youtube.com/watch?v=i7ft->

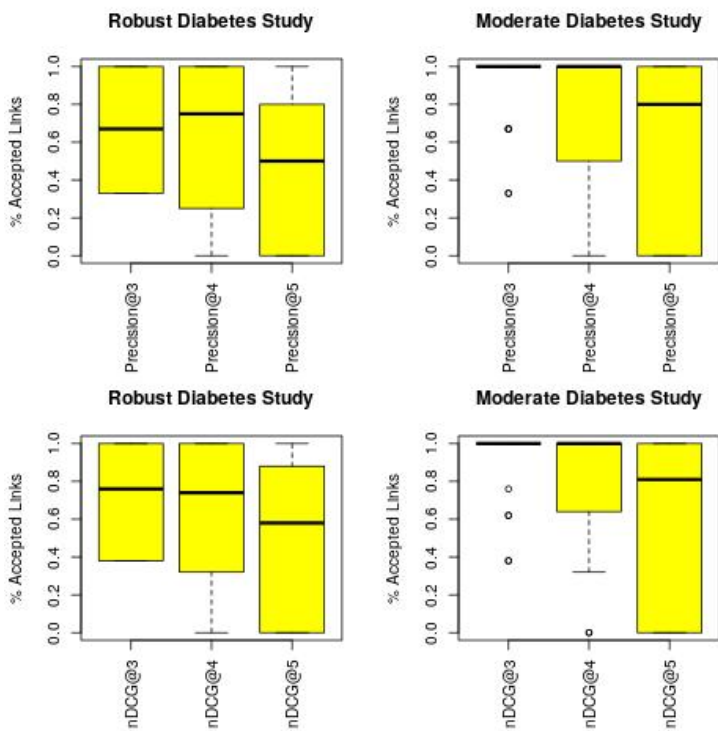


Figura 3.14: Resultados en Diabetes

nes. El algoritmo extrae el término *Inyección* (Shot) que no es referenciado en ninguna parte del contenido del vídeo. Así muestra este enlace recomendador y además es validado por los raters <sup>28</sup>. También el método corrige términos que fueron rechazados por no tener relación con el contenido del vídeo, por ejemplo términos como *esteroides* (steroid).

En la observación de Hipertensión/HTA se presentaron dificultades para llegar al acuerdo conforme a la sección 3.2.1. Entre otras cosas no se conseguía acuerdo mayor al coeficiente  $k > 0,8$ . La diferencia de los perfiles de los raters puede ser el motivo de estas dificultades, aún así se ha

6vR-Ic

<sup>28</sup>Childhood immunization <https://www.nlm.nih.gov/medlineplus/childhoodimmunization.htm>

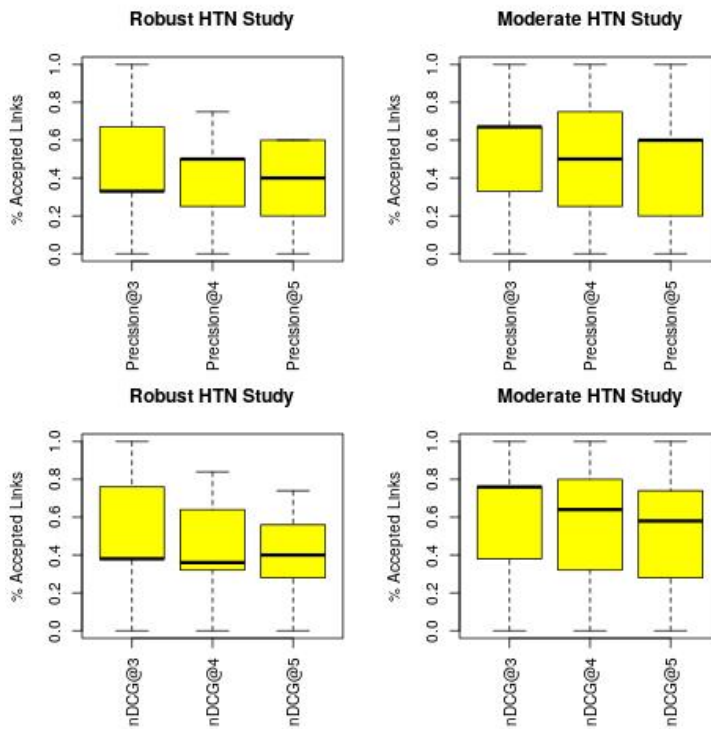


Figura 3.15: Resultados en Hipertensión/HTA

fijado un acuerdo con  $k > 0,6$ .

Con esta mejora se han previsto hasta la tercera y cuarta recomendación. A partir de una mayor precisión  $k > 4$ , conforme a la sección 3.2.2, el algoritmo decrece significativamente en su evaluación. La cuestión que se plantea es saber si hasta cuatro recomendaciones son suficientes para un paciente. En esta Tesis asumiremos que este es el caso, dado que existen fuentes que nos indican que muchas recomendaciones también pueden ser contraproducentes por saturar de información al paciente [Straus and Haynes, 2009, Oxman and Guyatt, 1993].

Será posible utilizar el algoritmo como una medida para recomendar contenido y vídeos para contrarrestar la información errónea o falsa. Por

---

**Algoritmo 4** función  $nDCG(x)$ 

---

**Entrada:**  $x$  total de recomendaciones a procesar**Salida:** Valor entre  $[0..1]$  correspondiente al resultado de la  $k$ -ésima recomendación

- 1:  $ideal = rep(1, length(x))$  //Crear un array de valores 1 de longitud el array de recomendaciones  $X$
  - 2:  $DCG = funcion(y) y[1] + suma(y[-1]/log(2:length(y), base = 2))$
  - 3: Devolver  $DCG(x)/DCG(ideal)$
- 

ejemplo, un vídeo que promueve el consumo de esteroides podría recomendar información alertando de los efectos potenciales negativos de su uso. Anticipamos que este algoritmo mejorará cuando haya más información disponible del perfil del usuario, lo cual abre un amplio abanico de posibilidades.

Esta fase dio solución a los objetivos planteados en la disertación correspondientes al punto 3b, 4, 5 y 6 de la sección 1.3.





Title: Juvenile Diabetes Research Foundation - Cure Video - Dallas  
Description

Since the founding of JDRF in 1970, the organization's largest and most successful fundraising program, the Walk to Cure Diabetes, has raised over 90 million dollars to go towards Type 1 diabetes research. Through this video the Greater Dallas Chapter hopes to bring more awareness to how diabetes affects those that are faced with the trials and tribulations of having the disease. More info at <http://www.jdrfdallas.org>

|                                   | Accept                | Reject                |
|-----------------------------------|-----------------------|-----------------------|
| <a href="#">Insulin-Dependent</a> | <input type="radio"/> | <input type="radio"/> |
| <a href="#">Diabetes Mellitus</a> | <input type="radio"/> | <input type="radio"/> |
| <a href="#">Diabetes Mellitus</a> | <input type="radio"/> | <input type="radio"/> |
| <a href="#">Shots</a>             | <input type="radio"/> | <input type="radio"/> |
| <a href="#">Diet</a>              | <input type="radio"/> | <input type="radio"/> |

Figura 3.16: Encuesta modelo de Precision@k Diabetes



# Capítulo 4

## Discussion

### 4.1. Related work

Broadly speaking, the main purpose of this research is the study of guidance systems recommendations in the area of health. Initially, the work focused on the use of graphs to manage recommendations. Some references [Li et al., 2012, Guan et al., 2009, Huang et al., 2004a, Huang et al., 2004b, Mirza, 2001], seemed to indicate that this was an attractive option, addressing recommendations in health systems. However, an in-depth analysis showed the great complexity of this approach, in particular in relation to how to explore and recreate the graph.

As an alternative, we raised the idea of exploiting “terms” to carry forward the study of recommender systems. Basically, the approach to the problem was: given a description of an item (e.g. a video), how should be selected the appropriate term from its description and how should it be injected into the system in order to give an useful recommendation to the user (patient).

Later on, it was observed that recommender systems could also help to filter out incorrect or harmful information (spam), following the lines already proposed in the literature [Burke et al., 2006, Zhang et al., 2006, Jindal and Liu, 2008, Ding et al., 2011, Steck, 2011, O’Donovan and Smyth, 2006].

As a first step, a review of the literature on health recommendations was performed, leading to a study of the state of the art in the chapter 2 which was published as a book chapter in [Sánchez-Bocanegra et al., 2015].

The next step, given the increase of usage of videos as a vehicle for the dissemination of health information, was to address the development of a system recommender of videos on health. Starting from Youtube<sup>TM</sup> as the most popular platform for the dissemination of videos [Davidson et al., 2010], and using the algorithm HealthTrust [Fernandez-Luque et al., 2012] to collect health videos from different channels and Playlists validated by the health community, the metadata associated with each video was analysed, including tags assigned by users. This seemed to be a promising path for video recommendation, as we found multiple works with tags and systems recommendations in other areas [Hurley, 2011, Givon, 2011, Sen et al., 2009b, Vig et al., 2010, Wray and Eklund, 2010, Donsbach et al., 2010, Vig et al., 2009, Shepitsen et al., 2008, Sen et al., 2009a]. The main problem arose when in 2012 Youtube<sup>TM</sup> blocked access to tags from their API, which led to a redefinition of the approach of our system.

How can we obtain “terms“ (excluding tags) that can be used for recommendations? The approach finally chosen for this thesis consisted of extracting health-related terms from the metadata of the title and the description of the video. These terms are then filtered with the clinical terminology SNOMED-CT, and from them related links from MedlinePlus (as a reliable source of health links) are recommended. To increase the number of recommendations, and to improve its quality, we used sanitary ontologies. (Bioontology) and natural language processing techniques (cTAKES) were proposed. Then these elements were tested in web applications that provided valid health information related to displayed videos, and finally an assessment of the accuracy of these recommendations was carried out. Two analyses were performed, one of them to determine the correlation between raters and the other one to validate the term itself. Table 3.7 shows the results of the first approximation, with a Kappa Cohen of 0,65 (good) in the studies of Diabetes and Hypertension, being in these cases the *precision@3* of 67% and 91% respectively. In contrast, later studies (as reflected in the tables 3.8 and 3.9) using ontologies and NLP, improve Kappa Cohen results up to good and very good for *precision@3* and *precision@4*, adding another statistic ( $nDCG_k$ ) to provide more guarantees of the algorithm. The results show that the increase in the text helps to grow up the recommendations. By applying natural language processing, we improved the quantity and quality of the recommendations (*precision@k* and  $nDCG_k$ ).

One of the main difficulties in this thesis was to anonymize user's profiles in pilot tests, which prevented us from using hybrids or collaborative recommendation methods. On the other hand this challenge allowed us to obtain recommendations independently from the users profile thanks to the use of ontologies and natural language processing.

As mentioned before our results show how the number of recommendations increase with the quantity of text analyzed. Rather than taking only the title of the video, to add description and subtitles improves makes the algorithm. Therefore, it could be thought that the automatic transcription of the audio that Youtube<sup>TM</sup> provides, might be an additional source of useful text. However, at the moment this transcription is not good enough, resulting in incomplete or incoherent texts.

#### 4.1.1. Milestones achieved in the thesis

##### **Is it possible to recommend reliable health content?**

Throughout the chapters 2 and 3 we have seen some possibilities. Also, literature review has proved that recommendations could help in raising awareness of the patient and improving knowledge about his illness This does not avoid the need for training and updating of health professionals as counselor and consultant in health.

##### **does ontologies help in the recommendation?**

One way to have reliable recommendations lies in adjusting and searching for clinical terms in dictionaries, validated by health professionals. Ontologies discriminate concrete health terms (or terminology).

##### **Does the recommender improve its performance by using NLP techniques?**

Natural language processing techniques constitute an element that, together with ontologies, improves the levels of analysis as it has been already described, which opens great possibilities in the improvement of the algorithm, especially when the user's profile is incorporated.

### **Is it better to use the API of SNOMED-CT rather than Bioontology API?**

Previous studies did shown that the algorithm was improved, it is true that more content (description) was included in the results. But Bioontology can open possibilities to other ontologies, a part from SNOMED-CT or MEDLINEPLUS, that can provide interesting results.

## **4.2. Future works**

The combination of the mapping of medical ontologies and natural language processing offers improvements in the accuracy of health recommendations, as we have been demonstrating with the exploitation of some ontologies (SNOMED-CT and MEDLINEPLUS) along with the exploration of the semantic content, This dissertation focuses on patient-oriented recommendations, but they could be extended to health stuff by using the remaining compendium of ontologies that exists in the Bioontology API,. Even more, it can be focused on any specialty within a health branch,ie., RXNorm (radiological Imaging), DDO (Diabetes), among others.

Another factor to consider is the option preserving users' anonymity, and to integrate recommender system in an environment with a health record or electronic health record. This would allow to adapt the recommendations to the user profile, needs or preferences. In addition, content from Social Media can be a source of useful information related to feelings or personal status of the user. Natural language processing tools integrate methods of analysis of feelings that would allow the management of recommendations (see section 2.5.3). Also automatic learning techniques (machine learning) may be included in the recommender algorithm.

The aim of this research was not to determine the performance improvements compared with traditional methods or other lines of recommendation. A more exhaustive study on the improvement of precision and seek time of the methods based on ontologies, natural language processing and automatic learning, could potentially be interesting.

Also the role of the user interfaces (UI/UX) referred to in section 2.5.1 can be essential. its study and evaluation would determine in which degree could be improve the recommendations and could be the subject of future

research.

A limitation of our research the modifications in the Youtube algorithms which are not transparent and can not be controlled. This also diculted the stability of our 3rd party software.

Finally, we did not test our recommender system with health record which would been the main potential of our research. if we include the users' profile, with his health medical record within a recommendation system, this will be significantly improved thanks to the use of the semantic - through the mapping of health ontologies - knowledge of users - with natural language processing. Therefore this will allow a more efficient and reliable approach using a hybrid recommendation system.

Another potential future work would be to include related scientific articles and clinical studies to the recommendations to support the decision making of healthcare professionals. For this purpose, we will use HealthRecSys engine and health ontologies (SNOMED-CT and MeSH) in order to provide a wide scope of scientific terms to facilitate the search on health scientific sources such as PubMed and ClinicalTrials.





# Apéndice A

## cTAKES configuration

```
<?xml version="1.0" encoding="UTF-8"?>
<cpedescription>
  <collectionReader>
    <collectionIterator>
      <descriptor>
<import location="FilesInDirectoryCollectionReader.xml"/>
      </descriptor>
      <configurationParameterSettings>
        <nameValuePair>
          <name>InputDirectory</name>
          <value>
<string>/caption-src/Diabetes/SRT</string>
          </value>
        </nameValuePair>
        <nameValuePair>
          <name>Language</name>
          <value>
            <string>en</string>
          </value>
        </nameValuePair>
        <nameValuePair>
          <name>Extensions</name>
```

```

        <value>
            <array>
                <string>srt</string>
                <string>txt</string>
            </array>
        </value>
    </nameValuePair>
</configurationParameterSettings>
</collectionIterator>
</collectionReader>
<collectionReader>
    <collectionIterator>
        <descriptor>
<import location="FilesInDirectoryCollectionReader.xml"/>
        </descriptor>
        <configurationParameterSettings>
            <nameValuePair>
                <name>InputDirectory</name>
                <value>
<string>/caption-src/Diabetes/SRT</string>
                </value>
            </nameValuePair>
            <nameValuePair>
                <name>Language</name>
                <value>
                    <string>en</string>
                </value>
            </nameValuePair>
            <nameValuePair>
                <name>Extensions</name>
                <value>
                    <array>
                        <string>srt</string>
                        <string>txt</string>
                    </array>
                </value>

```

```

        </nameValuePair>
    </configurationParameterSettings>
</collectionIterator>
</collectionReader>
    <casProcessors casPoolSize="3" processingUnitThreadCount="1">
<casProcessor deployment="integrated"
    name="AggregatePlaintextFastUMLSProcessor">
    <descriptor>
<import location="AggregatePlaintextFastUMLSProcessor.xml"/>
    </descriptor>
    <deploymentParameters/>
    <errorHandling>
        <errorRateThreshold action="terminate" value="0/1000"/>
        <maxConsecutiveRestarts action="terminate" value="30"/>
        <timeout max="100000" default="-1"/>
    </errorHandling>
    <checkpoint batch="10000" time="1000ms"/>
</casProcessor>
    <casProcessor deployment="integrated" name="Write CAS to XML file">
    <descriptor>
<import location="/cas_consumer/FileWriterCasConsumer.xml"/>
    </descriptor>
    <deploymentParameters/>
    <errorHandling>
        <errorRateThreshold action="terminate" value="0/1000"/>
        <maxConsecutiveRestarts action="terminate" value="30"/>
        <timeout max="100000" default="-1"/>
    </errorHandling>
    <checkpoint batch="10000" time="1000ms"/>
    <configurationParameterSettings>
        <nameValuePair>
            <name>outputDir</name>
            <value>
                <string>/caption-src/Diabetes/XML</string>
            </value>
        </nameValuePair>

```

```
        </configurationParameterSettings>  
    </casProcessor>  
</casProcessors>  
</cpedescription>
```

# Apéndice B

## Introduction on Health Recommender Systems chapter

Chapter 9 from *Methods in Molecular Biology* (Clifton, N.J.). 1246: 131-46. 2015.

## Introduction on Health Recommender Systems

C.L. Sanchez-Bocanegra, F. Sanchez-Laguna, and J.L. Sevillano

### Abstract

People are looking for appropriate health information which they are concerned about. The Internet is a great resource of this kind of information, but we have to be careful if we don't want to get harmful info. Health recommender systems are becoming a new wave for apt health information as systems suggest the best data according to the patients' needs.

The main goals of health recommender systems are to retrieve trusted health information from the Internet, to analyse which is suitable for the user profile and select the best that can be recommended, to adapt their selection methods according to the knowledge domain and to learn from the best recommendations.

A brief definition of recommender systems will be given and an explanation of how are they incorporated in the health sector. A description of the main elementary recommender methods as well as their most important problems will also be made. And, to finish, the state of the art will be described.

**Key words** Health recommender systems, Information research, Health information, Recommenders, Web 2.0

---

## 1 Introduction

Information and communication technologies (ICT) provide new ways of searching and gathering health information. Health consumers have access to a vast amount of different kinds of resources which are disseminated through the Word Wide Web [1].

Health content creators are continuously overloading the Internet with information. This makes the search for trusted health information more complicated, yet necessary.

People are demanding accurate and trustworthy health information. Search engines are in charge of this task and many studies analyze how these search engines determine health information to be trustworthy [2].

What users demand is trusted information selected according to their user profiles. Therefore, the so-called health recommenders find trustworthy health information and adapt it to the user profile, which can be obtained from their personal health record [3], in a process that contributes the empowerment of the patient [4–6].

Recommender methods depend mainly on user profiles, health information (also known as items) to search, and domains. Health recommender systems experience several difficulties [7] mainly cold-start, serendipity, sparsity, and spam (all these concepts will be explained in the following sections).

Furthermore, recommender systems still require improvements to produce more effective choices. In the next section we will present a brief description of health recommender systems.

---

## 2 Recommender Elements

Domains, user profile, and items are the main elements of a recommender system.

First, we describe a domain as the environment where all elements (items, users and their relationships) interact. For instance, if we are talking about diabetes, all documents, research, and relationships between clinicians and patients belong to the diabetes domain. Recommender methods usually fit best in a particular domain and may be useless in another; they are mostly focussed on and built for specific domains.

We can define a user profile as all those properties that identify a unique person within several domains. For example, a 28-year-old patient from Ireland suffering type 1 diabetes who measures his blood sugar levels and obtains 11.2 mmol/L. The recommender gathers all this information (place of birth, date of birth, diabetes type, blood sugar level) and includes it in the system. All of these properties determine one recommended choice or another.

Items are those elements that users are searching according to their needs. In many systems these items are documents, but generally an item is just a piece of information (for instance, drugs indications, dosing, use instructions, videos describing a disease).

Some recommenders collect those health information choices, process them all, and improve future recommendations in order to offer the best one. These interactions represent the essence of recommendations [8].

---

## 3 Basic Methods on Recommender Systems

### 3.1 Collaborative Approach

One of the two main options in recommender systems is the so-called collaborative approach, focussed on the users' social networks and the items they have selected in the past. The recommender processes them all in order to make future proposals. This approach analyzes the users' interactions in the past and shapes a recommendation. This model is also known as memory-based approach.

**Table 1**  
**User/item interaction matrix**

| User/item | Item 1 | Item 2 | ...                                   | Item $N$ |
|-----------|--------|--------|---------------------------------------|----------|
| User 1    |        |        |                                       |          |
| User 2    |        |        |                                       |          |
| ...       |        |        | (a) Integer value<br>(b) Like/dislike |          |
| User $M$  |        |        |                                       |          |

**Table 2**  
**Recommended items**

| Recommender items | Item 3 | Item 2 | ... | item $N$ |
|-------------------|--------|--------|-----|----------|
|                   |        |        |     |          |

In this approach, a table similar to Table 1 may be used to represent users in rows and items in columns. The intersections are filled with an integer or Boolean value. This helps us to determine the behavior among users.

We can also represent it as a matrix of items recommended for users (*see* Table 2).

The basic consideration of collaborative recommendation is “if users shared the same interests in the past, they would have similar tastes” [8] (*see* Fig. 1). The Collaborative Filtering (CF) answers these questions: How do users find others with similar taste? What about new users? What about new items?

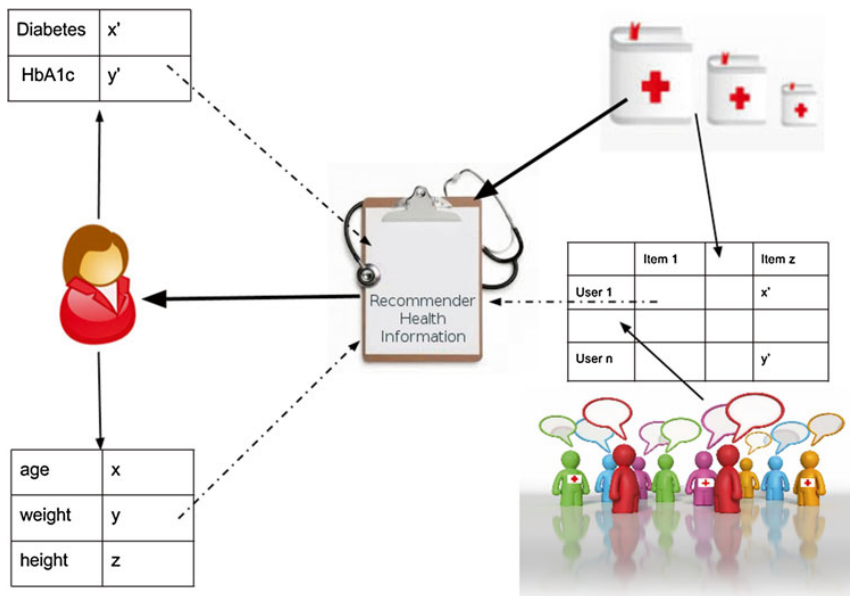
In this method, the rating of an item by a user is computed basically as a weighted sum of the ratings of other users. The weight in this sum is a measure of the similarity of the users, and there are several ways to compute it [7].

Pearson’s correlation is one of these ways to measure the users’ nearness. But this simple method is usually not enough to achieve the optimal recommendation of the recommendation; that is why it is used only in a few subsets of domains.

An alternative is the Cosine similarity measure, where users are considered as vectors in a  $n$ -dimensional space (with  $n$  being the number of items). Then we define a standard metric that measures the angle between them. The result may vary between 0 (orthogonal vectors, that is, no similarity) and 1 (equal) [7].

Both methods, Pearson’s Correlation and Cosine similarity, can be improved if we subtract the average rating behavior of the user. This way the fact that different users may rate differently is taken into account. The resulting method is called Adjusted Cosine





**Fig. 1** Collaborative approach

measure. Other options include Spearman's rank correlation coefficient and mean square difference measure [8].

Other more elaborate solutions found in the literature are inverse user frequency, significance weighting, and case amplification [7]:

- Inverse user frequency: Reduce the relative importance of those cases that receive universal agreement (such cases are rated more frequently but their rates are not very useful).
- Significance weighting: Two users may be highly correlated based on too few items, so a linear reduction of the similarity weight can be used in these cases. If a minimum of, say, 50 co-rated items is imposed, the prediction improvement is significant.
- Case amplification: Gives more weight to highly similar users through an amplification factor.

A different method is to try to reduce the computational complexity. For instance, when the nearest active neighbor with positive correlations is selected. After several selections, this method tends to select the same recommended items. However, it suffers from sparsity (insufficient available data leading to poor recommendations), so a surrogate method can be used which selects the  $K$  nearest neighbors (using the  $K$  nearest neighbor algorithm) [9, 10]. Both methods decrease the quality of the choice.

Many of above mentioned methods (correlation, cosine) can be adapted to compute similarities between items instead of users, and it has been shown that this approach may provide advantages in terms of computational performance [7].

### **3.2 Item-Based Nearest Neighbor Recommendation**

Defines the similarity between items. There is an offline processing that reduces subsequent real-time recommendations.

First we have to consider users as vectors in a  $n$ -dimensional space. Then we define a standard metric that measures the angle between them. The result may vary between 0 (no similarity) and 1 (equality) [7].

A better adaptation is the Adjusted Cosine measures that considers the average rating behavior of the user [10, 11].

Processing the entire matrix of users and items relations in real time could consume a lot of time and resources. To prevent that, all this data may be previously processed to obtain an item ratings' matrix for the community. This preprocessing is a usual approach in "learning machines." However, obtaining this matrix is very difficult because of the required minimum number of rates on the same item by different users, as well as the limited number of neighbors [12].

Finally, it is worth mentioning the "Slope One" method, which precomputes the average difference between the ratings of different users on one item [13]. This approach is easier to implement.

### **3.3 Content-Based Approach**

This approach considers item properties and user profiles as the essence of the recommendation (*see* Fig. 2). The idea is that if we know more about the item and/or the user the recommendation will be more accurate, even if the number of previous recommendations or the size of the user community is small. Many authors recognize as the model-based approach [8].

Each item may be represented by a list of properties (*see* Table 3) which contains all main characteristics to be identified.

First we need an off-line learning-mode phase. Only learned models are used to make recommendations online, so the system needs regular offline phases in order to improve its algorithms and make better recommendations each time [14].

Content-based approach looks for items similar to those that the user liked in the past, matching item attributes with the user profile to fit the best choice. Sometimes the entries in Table 3 are not meta-data, but simply keywords or terms appearing in the item (for instance, a document). In this case, content-based methods usually get documents (or text descriptions of other types of items) and simply filter the occurrences of these keywords. In the same manner this idea we can find the following basic methods.

### **3.4 The Vector Space Model and TF-IDF**

The vector space method analyses all terms that appear in an item, and encode items as vectors with the dimension of the number of terms. Two measures are used: Term Frequency (TF, or how often

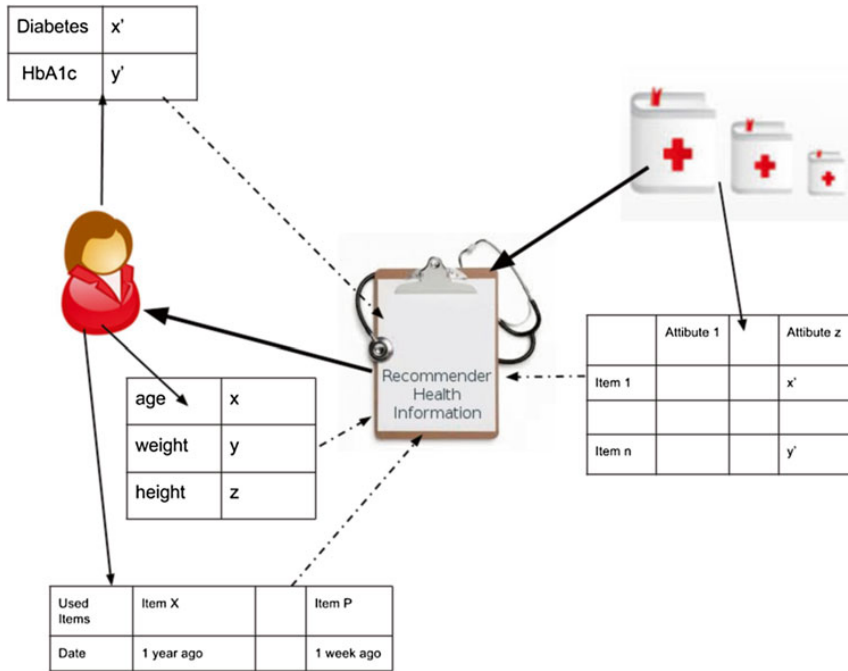


Fig. 2 Content-based approach

**Table 3**  
Items with explicit properties

| Item     | Size | Height    | Price |
|----------|------|-----------|-------|
| Item 1   | $X$  | $\gamma$  | $Z$   |
| Item 2   | $A$  | $B$       | $C$   |
| ...      | ...  | ...       | ...   |
| Item $X$ | $X'$ | $\gamma'$ | $Z'$  |

the term appears in the item) and Inverse Document Frequency (IDF, a measure intended to reduce the weight of terms that appear too often, which are not very useful) [8]. Note that although originally this method was used with keywords (terms) in documents (items), it could be used in other contexts (for instance, with movies if associated textual contents are available). There are several improvements of this method, mainly in the sense of reducing the size of these vectors and/or the required amount of information such as stopping words and stemming, size cutoff, phrases, and context [8].

Similarity-based retrieval is based on two main measures: item similarities and user likes/dislikes on previous items. Then, if a given number of similar items were liked before, a new item will be recommended. It uses k-nearest neighbor (kNN) algorithm and is simple to implement [15], but the prediction accuracy is lower than that of more complex methods.

The Rocchio method basically allows a user to rate the items (documents), and these ratings are incorporated into the user's profile. This feedback is used by the system to improve the query [16].

Probabilistic methods use an approach similar to that of a classification task, labeling the items according to previous user's ratings. Simple classification algorithms such as Naive Bayes classifier have been successfully used [8].

Machine learning is created to separate relevant and non-relevant items. The model may be based on machine learning techniques such as clustering, decision trees, neural networks, etc. One example is the Widrow-Hoff algorithm [17].

### 3.5 Knowledge-Based Approach

This method is a subset of the content-based approach. It creates a knowledge base of the items characteristics in order to improve the recommendation (see Fig. 3). In some cases, it can solve content-based approach weaknesses, for instance knowledge-based

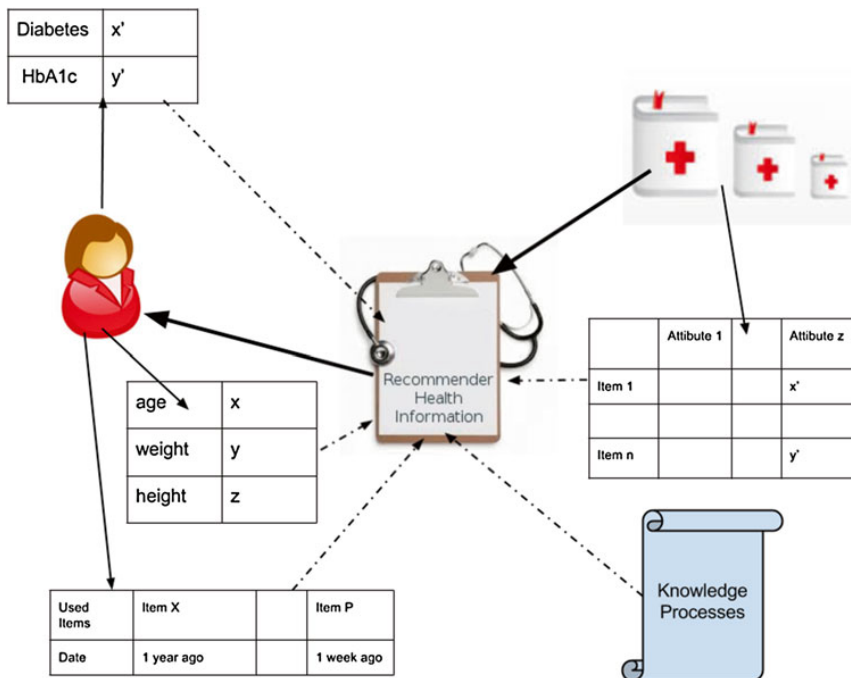


Fig. 3 Knowledge approach

recommendation may be used even with few available or useful ratings, as it can make recommendations based on the knowledge about the user and or/the item and not on previous explicit ratings.

Knowledge about the users can be obtained by asking them to specify their requirements. If the outcomes of the recommender are not suitable, the user may be asked to change the requirements. There are two basic types of knowledge-based recommenders: constraint based and case based. Constraint-based recommenders search for a set of items that fulfil the recommendation rules (or requirements). The process is similar to that of a Constraint Satisfaction Problem—CSP, that is, user's requirements and items characteristics form a set of variables and the recommender tries to obtain a solution for all these variables [8]. On the other hand, case-based recommenders try to find items similar to the user's requirements. To check this similarity a complete description (set of features) of the items is required [18].

---

## 4 Further Recommender Models

Recommender systems are taking advantage of some advancements and better models, several of which we will discuss next.

### 4.1 Matrix Factorization/Latent Factor Models

One of the common problems with Information retrieval consists of finding data in documents. Given a user's query, a possible solution is to find data that coincides with that query, a solution that can be used in recommended systems. However, this method has a set of latent hidden factors (for instance, synonyms or polysemous words in textual documents) [8]. Possible solutions include:

- Singular Value Decomposition (SVD), which algebraically reduces the dimension of the associated matrices (formed by the users and items) [19].
- Latent Semantic Analysis (LSA) or latent semantic indexing (LSI), discovers the latent factors and reduces the size of these matrices by merging semantically similar elements, sometimes in conjunction with SVD [20, 21].
- LSI-based retrieval: make it possible to retrieve relevant documents even if it does not contain many words of the user's query [22].

### 4.2 Association Rule Mining

This method finds useful patterns in a transaction dataset. The association rules are following the form  $X \rightarrow Y$  where  $X$  and  $Y$  are two disjoint subsets of all available items that satisfy constraints on measures of the significance and interestingness [23]. This method can be used in health recommender systems. For instance, consider a user looking for information on diabetes. If he/she downloads a video explaining the possible complications of the disease, a recommender may offer him/her videos on healthy lifestyles.

**Table 4**  
**True/false transactions for each item**

| # Transactions | Item 1 | Item 2 | ... | Item $N$ |
|----------------|--------|--------|-----|----------|
| 1              | True   | False  |     | True     |
| 2              | False  | True   |     | False    |
| ...            |        |        |     |          |
| $n$            | True   | True   |     | False    |

This method is represented in an iteration item matrix; each intersection cell contains a boolean value: True means presence and False means absence (*see* Table 4).

The matrix contains almost thousands of transactions. The method finds rules that correlate the presence of one set of items with another set.

This method may be evaluated as successful if given  $X$ ,  $Y$  as items, it finds rules  $X$  and  $Y \rightarrow Z$  with minimum support (probability that a transaction contains the three items) and confidence (conditional probability that a transaction with  $X$  and  $Y$  also contains  $Z$ ) [24].

### 4.3 Ontologies and Semantic web Recommendations

The Resource Description Format (RDF) is a standard that represents information modeled as a “graph.” RDF is one of the pillars of the so-called Semantic Web, where it can be used to serialize information represented using graphs. Together with languages like OWL (Ontology Web Language), these developments allow an enhanced representation of the information which can be used to obtain improved recommendations [25, 26].

Ontologies support coding systems both in electronic healthcare records [27] or social media [28], gathering the precise meaning of one term, and determining relationships between terms. A term is made of words that represent an item. Either terms or relations help the system to improve the recommendations it makes [29].

### 4.4 Hybrid Methods

Hybrid recommender systems combine two or more recommendation approaches to gain better performance. Most commonly, collaborative filtering is combined with some other techniques (as content based or knowledge based) in order to minimise their respective weaknesses [30].

---

## 5 Challenges on Recommender Systems

In this section we briefly describe some of the challenges that research on recommender systems have to address.

### **5.1 Implicit and Explicit Ratings**

Some recommenders ask for explicit item ratings to obtain users' opinion more precisely (5-points, 7-points, like/dislike). Using these ratings allows more precise user recommendations (for example, a 10-points scale is better accepted in movies recommendations) though it requires additional effort from users [8].

On the other hand, implicit ratings collect external properties from the environment. For example, if someone buys food in the city of Málaga (southern Spain) in summer at 13 pm, the place and time are considered implicit ratings to provide healthy recommendations adequate for the usually hot and humid weather. Therefore, we cannot be sure whether this user behavior is correctly interpreted with these ratings.

Obtaining useful ratings without additional efforts from users is one of the challenges of recommender systems.

### **5.2 Data Sparsity**

Another typical problem of recommender systems, particularly important with those using a collaborative approach, is that data and ratings tend to be sparse. A good solution is to use user profile information when calculating user similarity [7]. That is, if two users suffer from the same disease they could be considered similar even if they have not rated the same item (video, document) similarly. Other characteristics like gender, age, education, and interests could help in classifying the user. Additionally, methods like matrix factorization or latent factor models can be used to reduce the size and dimension of the rating matrix [9].

### **5.3 Cold-Start**

It takes time to include new users and items that have not been rated yet, so several questions arise: How can recommendations be made to new users that have not rated yet? And how to deal with items that have not been rated yet either?

Some approaches avoid the new-user problem by asking the user for a minimum number of ratings before the service can be used. Others exploit the supposed "transitivity" between them through relationship graphs [8]. But the most usual approach is adopting hybrid methods, that is, a combination of content-based and collaborative methods [7].

### **5.4 Serendipity/Overspecialization**

Some recommenders (mainly content based) rely on the similarity of items, and the user's interest is taken for granted. However, in many contexts this assumption is wrong. For instance, when the system can only recommend items that score highly against a user's profile, the recommendation may be too obvious to be useful. This undesirable effect is called overspecialization. On the contrary, increasing the serendipity is usually useful, that is, sometimes the recommender can offer unsearched but maybe useful items, even with a certain randomness [31].

### 5.5 Latency

This problem appears on new items (mainly with the collaborative approach), the system is unable to select recently added items which need to be reviewed before they can be recommended [32].

---

## 6 Health Recommender Systems

Nowadays more than 80 % of the Internet users have searched for health information [33]. Patients can identify the reasons of their disease, find treatments, learn healthy habits or contact other people suffering from the same disease, and as a result they are continuously increasing their knowledge and empowerment.

But the amount of information is so huge that sometimes it is difficult for the users to find trusted health information. Identifying relevant resources is very difficult for the user; they may find confusing and misleading information about their disease, most of them with a very low quality. Additionally they want to access this information as quickly as possible. Therefore, the role of recommender systems is important, and required characteristics are usefulness, trustworthiness, performance, and ability to adapt the outcomes according to users' profiles and needs [34].

Collaborative recommendations were the most used recommenders on health sector [7]. However, health decisions depend on the knowledge of the patient's needs, and this knowledge is often incomplete. This results in errors or omissions and eventually adverse outcomes.

### 6.1 Personal Health Record and Recommender Systems

Personal health record (PHR) is a health record that contains information related to the care of a patient [35]. PHRs are usually maintained by the patient or a carer, so a useful approach is having a health recommender system include PHR as a part of the user profile [34]. Of course, there are many privacy issues related to this approach, but still it seems clear that a reduced version of the user's PHR, which carefully excludes the most sensible information, would be very useful to define the user's profile of a health recommender system.

All approaches discussed so far, collaborative, content based, and knowledge based, can be used together with information from personal health records to provide useful recommendations, with the limitations mentioned in Subheading 5. Sometimes hybrid models are used together with the personal health record as part of the user profile (*see* Fig. 4). Furthermore, mobile recommender systems are becoming more widespread as users increase the access to health information through mobile systems [36].

A way to exploit recommendation based on the personal health record is through semantic networks like Wikipedia [34]. A semantic network is a network that represents relations between concepts as a knowledge representation. Normally such a network is



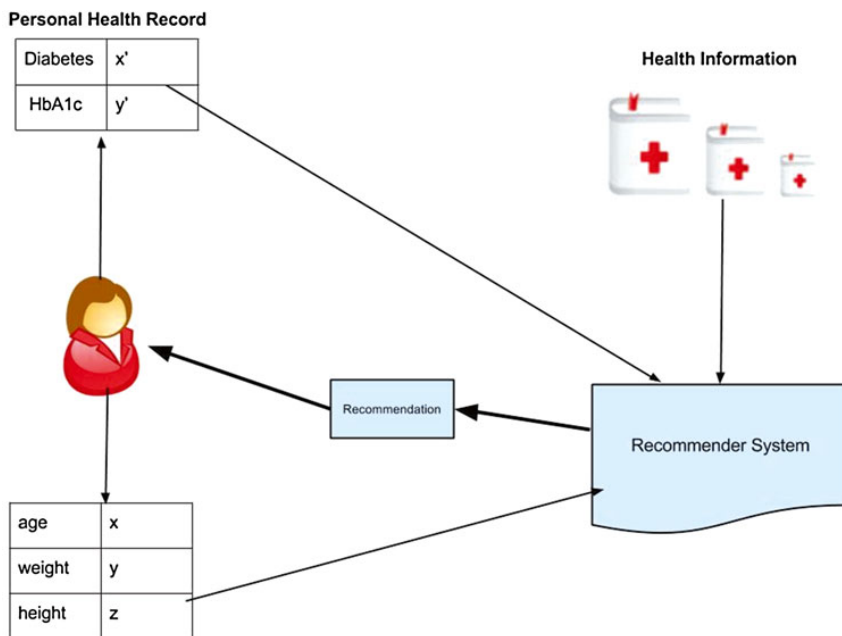


Fig. 4 Health recommender system

represented by a graph where each node represents the concept (subject) and the edges represent the relations (verb) [37].

Figure 5 represents an example of a semantic network. Nodes contain the health concept (subject and object) and the edges associated with these nodes represent their relations. This example, extracted from JMIR [38], includes a brief representation on SQL tables. The resulting recommender system uses the distances among nodes as its main criterion.

One recent concept related to the collaborative nature of recommended systems was introduced by Eysenbach, who named it “apomediation,” a “socio-technological term” that identifies “trustworthy and credible information and services” [39]. Apomediation refers to the process of guiding users to information or services through different “agents” that are not clearly identified as intermediaries; instead, they may be other users and/or automatic systems (recommenders) that filter the information collaboratively. As a result, the credibility and trustworthiness cannot be directly guaranteed. The recommender systems need to include a way to ensure trusted health information, fitting the information according to the user needs, which is the main challenge for health recommender systems.

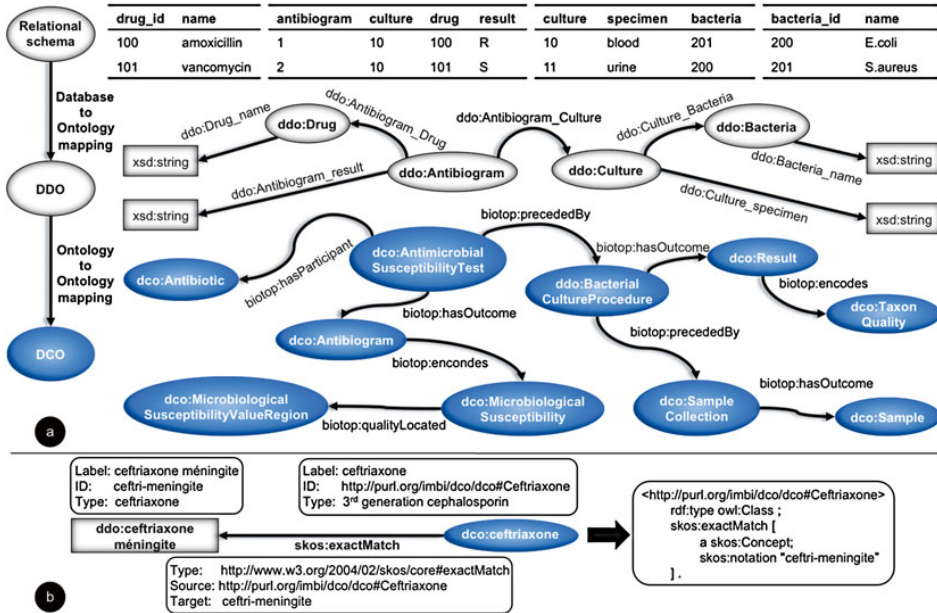


Fig. 5 Example of semantic network. Source: JMIR (Example of a semantic network: the hybrid ontology-driven interoperability mapping model. JMIR. <http://www.jmir.org/2012/3/e73/>) [38]

### 6.2 Social Media and Recommender Systems

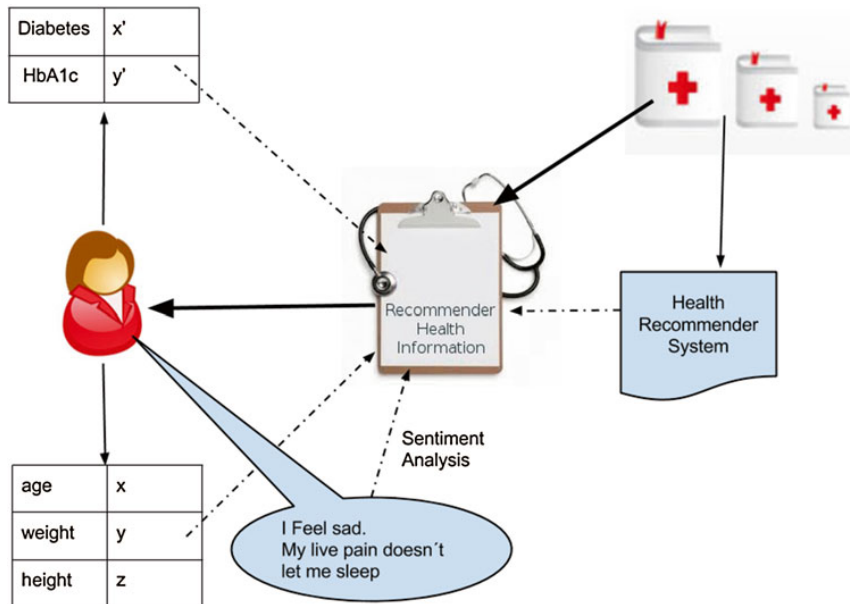
Health professionals and patients are taking advantage of social media, using the Internet as a tool to find and share health information. They can benefit from using the Internet to connect with each other, and to provide other helpful insights, but it is essential to be aware of possibly unwanted consequences.

One way to avoid these effects is to take advantage of health educational resources. Recommender systems and personalized health education work together to filter the overwhelming information and focus on the actual user needs [5]. The explanation-driven approach to a personalized content recommendation for patients is an example of this thread [40]. The information is accompanied with detailed explanations that educate patients to make informed medical decisions.

### 6.3 Sentiment Analysis

Machine learning increases the potential of recommender systems; the system learns with the interaction of the users and adapts its decisions according to previous choices. The understanding increases with the recommender process.

Machine learning is effective in sentiment classifications [41]. Sentiment analysis obtains subjective information from the context. People are continuously communicating their feelings; we can



**Fig. 6** Sentiment analysis and health recommender system

for instance look for any verbs expressing sentiments (to feel, to want, to love) or nouns (pain, loss, feelings) and try to adapt the recommendation following the sentiment context (*see* Fig. 6).

The treatment of opinions, sentiments, and subjective texts is one of the main research interest of current recommender systems [42]. Feelings are one of the implicit values to be included in content-based systems. Social networks (like Facebook, Twitter, Youtube) are examples where these messages can be retrieved to be processed on the health recommender system. Moreover there are other health dedicated channels that contain sentiment messages (like Patientlikeme,<sup>1</sup> Crochane,<sup>2</sup> Webicina<sup>3</sup>). Information from the abovementioned social media can be retrieved and linked with trusted health sources. This process increases the possibilities in recommenders.

<sup>1</sup>Web Patientlikeme <http://www.patientslikeme.com/>

<sup>2</sup>Web Chrocane <http://www.cochrane.org/>

<sup>3</sup>Web Webicina <http://www.webicina.com/>

## References

1. Fox S, Jones S (2009) The social life of health information. Pew Internet & American Life Project, Washington, DC, 2009-12
2. Fernandez-Luque L, Karlsen R, Melton GB (2012) Healthtrust: a social network approach for retrieving online health videos. *J Med Internet Res* 14(1):e22
3. Cline RJW, Haynes KM (2001) Consumer health information seeking on the Internet: the state of the art. *Health Educ Res* 16(6): 671–692
4. Khan SA, McFarlane DJ, Li J, Ancker JS, Hutchinson C, Cohall A et al (2007) Healthy Harlem: empowering health consumers through social networking, tailoring and web 2.0 technologies. *AMIA Annu Symp Proc* 2007 Oct 11:1007
5. Fernandez-Luque L, Karlsen R, Vognild LK (2009) Challenges and opportunities of using recommender systems for personalized health education. *MIE*, pp 903–907
6. Pattaraintakorn P, Zaverucha GM, Cercone N (2007) Web based health recommender system using rough sets, survival analysis and rule-based expert systems. In: An A, Stefanowski J, Ramanna S, Butz CJ, Pedrycz W, Wang G (eds). *Rough sets, fuzzy sets, data mining and granular computing* [Internet]. Berlin: Springer. Available from: [http://link.springer.com/chapter/10.1007/978-3-540-72530-5\\_59](http://link.springer.com/chapter/10.1007/978-3-540-72530-5_59)
7. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans Knowl Data Eng* 17(6):734–749
8. Zanker M, Felfernig A, Friedrich G (2011) *Recommender systems: an introduction*. Cambridge University Press, New York, NY
9. Sarwar B, Karypis G, Konstan J, Riedl J (2000) Application of dimensionality reduction in recommender system - a case study. *Proc. ACM workshop on web mining for e-commerce-challenges and opportunities -WebKDD, Boston MA*. ACM, New York, NY
10. Melville P, Sindhvani V (2010) Recommender systems. In: Sammut C, Webb G (eds) *Encyclopedia of machine learning*. Springer, New York, NY
11. Symeonidis P, Nanopoulos A, Papadopoulos AN, Manolopoulos Y (2008) Collaborative recommender systems: combining effectiveness and efficiency. *Exp Syst Appl* 34(4): 2995–3013
12. Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7(1): 76–80
13. Wang P, Ye H (2009) A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering. *International Conference on Industrial and Information Systems, 2009 IIS'09*, pp 152–154
14. Lops P, Gemmis M de, Semeraro G (2011) Content-based recommender systems: state of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds). *Recommender systems handbook* [Internet]. New York, NY: Springer. Available from: [http://link.springer.com/chapter/10.1007/978-0-387-85820-3\\_3](http://link.springer.com/chapter/10.1007/978-0-387-85820-3_3)
15. Suchal J, Návrat P (2010) Full text search engine as scalable k-nearest neighbor recommendation system. In: Bramer M (ed). *Artificial intelligence in theory and practice III* [Internet], Berlin: Springer, pp 165–73. Cited September 26 2013. Available from: [http://link.springer.com/chapter/10.1007/978-3-642-15286-3\\_16](http://link.springer.com/chapter/10.1007/978-3-642-15286-3_16)
16. Pazzani MJ, Billsus D (2007) Content-based recommendation systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds). *The adaptive web* [Internet]. Berlin: Springer, pp 325–341. Cited September 26 2013. Available from: [http://link.springer.com/chapter/10.1007/978-3-540-72079-9\\_10](http://link.springer.com/chapter/10.1007/978-3-540-72079-9_10)
17. Lewis DD, Schapire RE, Callan JP, Papka R (1996) Training algorithms for linear text classifiers. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* [Internet], New York, NY: ACM, pp 298–306. Cited September 26 2013. Available from: <http://doi.acm.org/10.1145/243199.243277>
18. Smyth B (2007) Case-based recommendation. In: Brusilovsky P, Kobsa A, Nejdl W (eds) *The adaptive web*, vol 4321, LNCS. Springer, Berlin, pp 342–376
19. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37
20. Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. *Disc Process* 25(2–3):259–284
21. Hofmann T (1999) Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* [Internet]. New York, NY: ACM, pp 50–57. Available from: <http://doi.acm.org/10.1145/312624.312649>
22. Dumais ST (1991) Improving the retrieval of information from external sources. *Behav Res Methods Instrum Comput* 23(2):229–236

23. Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. *SIGKDD Explor Newsl* 2(1):58–64
24. Sarawagi S, Thomas S, Agrawal R (1998) Integrating association rule mining with relational database systems: alternatives and implications. Proceedings of the 1998 ACM SIGMOD international conference on Management of data [Internet]. New York, NY: ACM, pp 343–354. Available from: <http://doi.acm.org/10.1145/276304.276335>
25. Carroll JJ, Dickinson I, Dollin C, Reynolds D, Seaborne A, Wilkinson K (2004) Jena: implementing the semantic web recommendations. Proceedings of the 13th international World Wide Web conference on alternate track papers & posters [Internet]. New York, NY: ACM, pp 74–83. Available from: <http://doi.acm.org/10.1145/1013367.1013381>
26. Tomanek K, Wermter J, Hahn U (2007) Efficient annotation with the Jena ANnotation Environment (JANE). Association for computational linguistics, pp 9–16. Available from: <http://dl.acm.org/citation.cfm?id=1642059.1642061>
27. Rector AL, Qamar R, Marley T (2009) Binding ontologies and coding systems to electronic health records and messages. *Appl Ontol* 4(1):51–69
28. Bedi P, Kaur H, Marwaha S (2007) Trust based recommender system for the semantic web. Proceedings of the 20th international joint conference on artificial intelligence [Internet]. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp 2677–2682. Cited November 14 2013. Available from: <http://dl.acm.org/citation.cfm?id=1625275.1625706>
29. Peis E, Morales-del-Castillo JM, Delgado-López JA (2008) Analysis of the state of the topic [Internet]. “Hiptertext.net”, num. 6
30. Burke R (2002) Hybrid recommender systems: survey and experiments. *User Model User Adap Inter* 12(4):331–370
31. Iaquinta L, de Gemmis M, Lops P, Semeraro G, Filannino M, Molino P (2008) Introducing serendipity in a content-based recommender system. Eighth international conference on hybrid intelligent systems, 2008 HIS’08, pp 168–173
32. Sollenborn M, Funk P (2002) Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems. In: Craw S, Preece A (eds). *Advances in case-based reasoning* [Internet]. Berlin: Springer, pp 395–405. Cited October 17 2013. Available from: [http://link.springer.com/chapter/10.1007/3-540-46119-1\\_29](http://link.springer.com/chapter/10.1007/3-540-46119-1_29)
33. Pew Internet & American Life Project (2013) Health topics—use of internet users look for health information online. <http://pewinternet.org/Reports/2013/Health-online.aspx>
34. Wiesner M, Pfeifer D (2010) Adapting recommender systems to the requirements of personal health record systems. Proceedings of the 1st ACM international health informatics symposium [Internet]. New York, NY: ACM, pp 410–414. Cited October 25 2013. Available from: <http://doi.acm.org/10.1145/1882992.1883053>
35. Kimmel Z, Greenes RA, Liederman E (2004) Personal health records. *J Med Pract Manage* 21(3):147–152
36. Ricci F (2010) Mobile recommender systems. *Inf Tech Tour* 12(3):205–231
37. Spitzer M, Braun U, Hermle L, Maier S (1993) Associative semantic network dysfunction in thought-disordered schizophrenic patients: direct evidence from indirect semantic priming. *Biol Psychiatry* 34(12):864–877
38. Teodoro D, Pasche E, Gobeill J, Emonet S, Ruch P, Lovis C (2012) Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation. *J Med Internet Res* 14(3):e73
39. Eysenbach G (2008) Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res* 10(3):e22
40. Roitman H, Messika Y, Tsimmerman Y, Maman Y (2010) Increasing patient safety using explanation-driven personalized content recommendation. Proceedings of the 1st ACM international health informatics symposium [Internet]. New York, NY: ACM, pp 430–434. Cited September 22 2013. Available from: <http://doi.acm.org/10.1145/1882992.1883057>
41. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol 10 [Internet]. Stroudsburg, PA: Association for Computational Linguistics, pp 79–86. Cited October 29 2013. Available from: <http://dx.doi.org/10.3115/1118693.1118704>
42. Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2(1–2):1–135



# Apéndice C

## HealthRecSys paper

HealthRecSys: A semantic content-based recommender system to complement health videos Paper for BMC Medical Informatics and Decision MakingPaper.

RESEARCH ARTICLE

Open Access



# HealthRecSys: A semantic content-based recommender system to complement health videos

Carlos Luis Sanchez Bocanegra<sup>1</sup>, Jose Luis Sevillano Ramos<sup>1</sup>, Carlos Rizo<sup>2</sup>, Anton Civit<sup>1</sup> and Luis Fernandez-Luque<sup>3\*</sup> 

## Abstract

**Background:** The Internet, and its popularity, continues to grow at an unprecedented pace. Watching videos online is very popular; it is estimated that 500 h of video are uploaded onto YouTube, a video-sharing service, every minute and that, by 2019, video formats will comprise more than 80% of Internet traffic. Health-related videos are very popular on YouTube, but their quality is always a matter of concern. One approach to enhancing the quality of online videos is to provide additional educational health content, such as websites, to support health consumers. This study investigates the feasibility of building a content-based recommender system that links health consumers to reputable health educational websites from MedlinePlus for a given health video from YouTube.

**Methods:** The dataset for this study includes a collection of health-related videos and their available metadata. Semantic technologies (such as SNOMED-CT and Bio-ontology) were used to recommend health websites from MedlinePlus. A total of 26 health professionals participated in evaluating 253 recommended links for a total of 53 videos about general health, hypertension, or diabetes. The relevance of the recommended health websites from MedlinePlus to the videos was measured using information retrieval metrics such as the normalized discounted cumulative gain and precision at K.

**Results:** The majority of websites recommended by our system for health videos were relevant, based on ratings by health professionals. The normalized discounted cumulative gain was between 46% and 90% for the different topics.

**Conclusions:** Our study demonstrates the feasibility of using a semantic content-based recommender system to enrich YouTube health videos. Evaluation with end-users, in addition to healthcare professionals, will be required to identify the acceptance of these recommendations in a nonsimulated information-seeking context.

**Keywords:** Patient Education, Health Recommender System, Natural Language Processing, Information Retrieval

## Background

Recent studies have shown an increasing trend in the use of the Internet as a search tool for health-related information [1–3]. Web 2.0 [4] allows contributions from any user in a network, which has given rise to a wealth of health-related information with a wide range of co-existing trustworthy sources [5, 6]. For this reason, screening tools can assist users in selecting relevant information.

Recommender systems are among the many solutions used to obtain valid information. When searching for an item, users obtain a list of recommended results that may match their preferences. Various filtering methods make it possible to refine and tailor these recommendations [7, 8]. Recommender systems can be divided into three basic groups: collaborative, context-based, and hybrid systems. Collaborative systems build on experience gathered from previous user experiences, i.e., items previously chosen by other users shape future results [9]. Context-based systems focus on the characteristics of an item, i.e., when searching for a camera, the recommendation output is based on its resolution, price, and color. Hybrid

\* Correspondence: lluque@qf.org.qa

<sup>3</sup>Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar Foundation, PO Box 5825, Doha, Qatar  
Full list of author information is available at the end of the article





recommender systems combine features of context-based and collaborative systems [10]. Recommender systems can be also used to give additional item recommendations for a given item, such as the related videos that are shown by YouTube next to the user's current video. These recommendations often relay user ratings, but can also be based on knowledge-based systems.

Recommender systems have been used in several applications for finding accurate information. They were introduced as a computer-based intelligent technique that assists people with the problem of information overload. These systems provide personalized solutions in various specific domains [11–13]. Recommender systems reflect the user's interest and make proper personalized recommendation through several methods. Most current systems have adopted recently developed algorithms that use machine-learning [14–16], naive Bayes [16, 17], social-trust-based [18–21], constraint-based [22], case-based [23, 24], and matrix factorization [25, 26] approaches. Recommender systems are also found in clinical settings, mainly to assist health professionals, though some systems assist family members, patients, or caregivers [27–29].

Recent advancements in online recommender systems are enhanced by the “Semantic Web” [30], which allows for the extraction of vast amounts of information through metadata mining and artificial intelligence techniques [31]. Using these techniques, it is possible to rank and classify items based on terms that encompass several properties grouped into ontologies [32]. In the life sciences, ontologies play an important role in filtering relevant item and creating knowledge-based systems. Knowledge-based, case-based, and social-trust-based approaches utilize user metadata, such as age and gender, to define recommendation rules. Machine-learning and naïve Bayes methods create models to learn users' interests from their historical behavior. Matrix factorization learns a user's latest interests by collaboratively factoring the rating matrix over historically recorded user-item preferences.

Health terms are also grouped into ontologies, creating an important potential resource for many applications, including recommender systems. Health ontologies usually have an application-programming interface (API) to precisely define their operation. One example of an API<sup>1</sup> is Bio-ontology,<sup>2</sup> which contains more than 600 health-related ontologies. Using Bio-ontology, Rivero-Rodriguez et al. recommended relevant links for a subset of health-related YouTube videos [33] by extracting corresponding clinical terms from the Medline Plus API for the International Health Terminology Standards Development Organization, which maintains SNOMED-CT, a multilingual clinical healthcare ontology.<sup>3</sup>

## Our previous work

This study is based on our previous work. Fernandez-Luque et al. reused algorithms from [33], but added the Bio-ontology API to improve the results for obtaining links from Medline Plus. In this study, we also rely on diabetes videos [34] for which we have already explored the use of semantic technologies to provide additional content recommendations [35]. Based on [33, 34], the proposed method gathers recommendations for Medline Plus links (see Fig. 1) from video subtitles to increase the number of associated terms using health ontologies. An additional movie file shows this in more detail [see Additional file 1]. An important limitation, both in the current and previous recommender systems, stems from the difficulty of mapping suitable terms to the ontology, especially when extracting representative terms from video content. One interesting approach to this problem uses natural language processing (NLP) [36–38] techniques, which can combine syntactic, semantic, and contextual analyses. NLP has previously been used in healthcare [39, 40], especially for mining electronic health records [41].

## Objectives

In online browsing, it is common to search for content related to online material currently being viewed. For example, after watching a video on YouTube, the watcher might look for additional content as part of an information seeking strategy. This search strategy has led to the creation of recommender systems that provide recommendations for related content. In this study, we explore the feasibility of recommending links to health educational content as a supplement to online health videos, focusing on recommendation methods that use semantic-based technologies to enhance online health content recommender systems. Further, this study investigates website recommendations that will enhance health videos, because video formats have shown the fastest growth on the Internet and it is estimated that, by 2019, video will constitute more than 80% of Internet traffic.<sup>4</sup>

## Methods

In this study, we introduce HealthRecSys, a recommender system with Bio-ontology terms that generates Medline Plus links from text extracted from the metadata of selected YouTube videos (see Fig. 1).

Our recommender system involves several steps: A) collecting filtered words from the title of a video, B) collecting any one SNOMED-CT term from the title, C) collecting a group of SNOMED-CT terms from the title, and D) determining the union of the results of steps B and C. Step A uses a “stop word” filtering system (i.e., that avoids preposition, adverbs, and similar terms), and

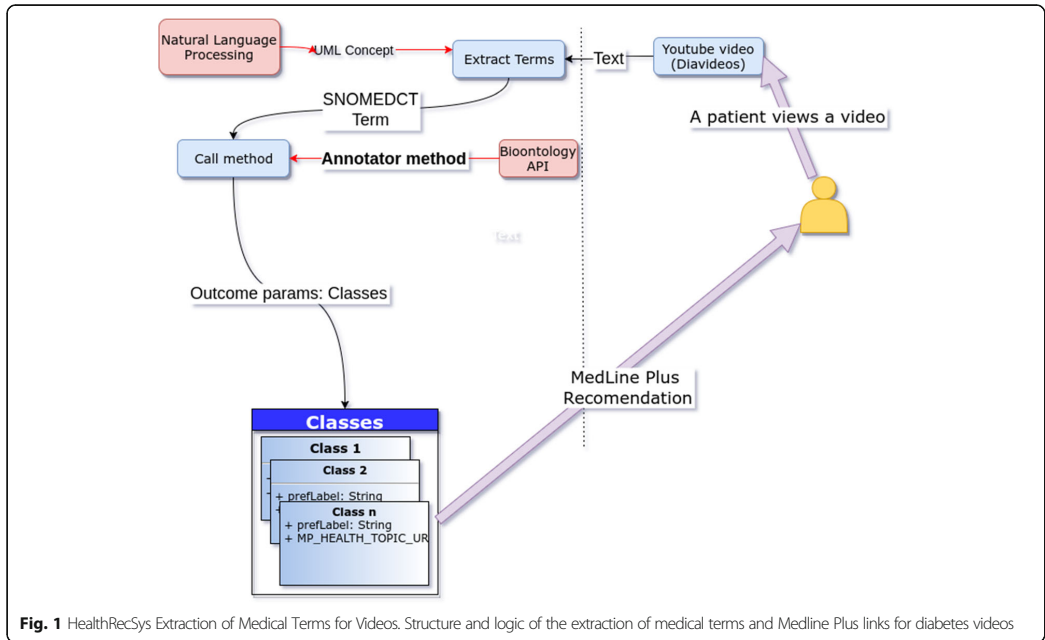


Fig. 1 HealthRecSys Extraction of Medical Terms for Videos. Structure and logic of the extraction of medical terms and Medline Plus links for diabetes videos

steps B and C are combined with SNOMED-CT to extract web links from MedlinePlus.

**Algorithm design**

We selected keywords (or terms) from video metadata (i.e., the video title, description, and subtitles). These keywords are used to identify semantic terms from Medline Plus. Fig. 1 shows the **Term Extraction** process for diabetes videos from YouTube. The algorithm contains two steps:

- **Source term collection:** the video title, description, and subtitle are collected as possible terms.
- **NLP:** In this case, NLP is applied to the title description and video subtitles using the cTAKES

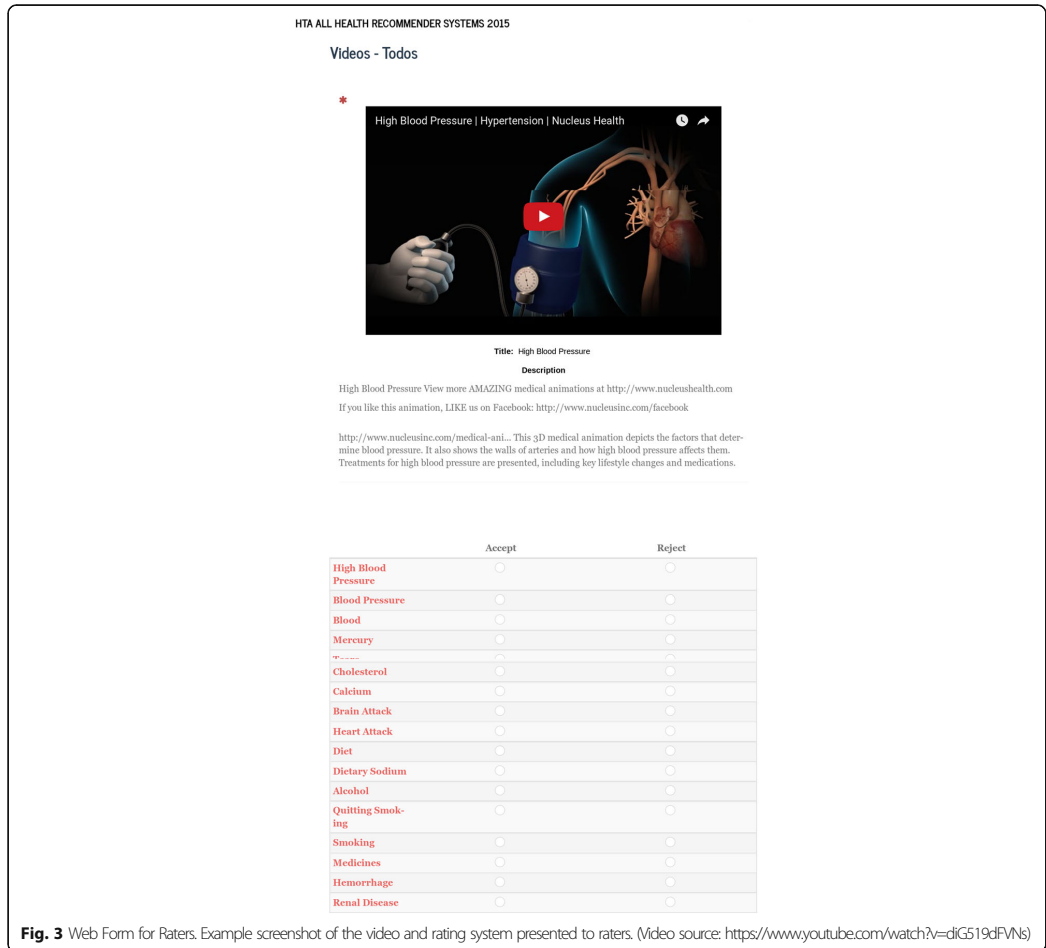
framework.<sup>5</sup> This is a health-specific NLP implementation that extracts SNOMED-CT health terms from text. See Fig. 2 for an example of extracted metadata of a video.

We conducted a text analysis using the Unified Medical Language System (UMLS)<sup>6</sup> with SNOMED-CT annotations to match the cTAKES framework. To achieve this, we inject the original video metadata files (with title, description, and subtitle) procedures from the UMLS library, resulting in an XML file that contains a morphological, syntactic, and semantic analysis.

From this file, we filtered the *UmlConcept* labels that contain collected terms from the SNOMED-CT ontology properties. For instance, Fig. 2 shows example XML for the terms *Blood*, *Entire Cell*, and *Cells*. The cTAKES

```
<org.apache.ctakes.typesystem.type.refsem.UmlConcept_id="24523" codingScheme="SNOMEDCT" code="87612001" oid="87612001#SNOMEDCT"
score="0.0" disambiguated="false" cui="C0005767" tui="T024" preferredText="Blood" >
- <uima.cas.FSArray_id="24533" size="1">
  <i>24523</i>
  <uima.cas.FSArray>
  <org.apache.ctakes.typesystem.type.refsem.UmlConcept_id="24488" codingScheme="SNOMEDCT" code="362837007" oid="362837007#SNOMEDCT"
score="0.0" disambiguated="false" cui="C1269647" tui="T025" preferredText="Entire cell" >
- <org.apache.ctakes.typesystem.type.refsem.UmlConcept_id="24478" codingScheme="SNOMEDCT" code="4421005" oid="4421005#SNOMEDCT"
score="0.0" disambiguated="false" cui="C0007634" tui="T025" preferredText="Cells" >
- <uima.cas.FSArray_id="24498" size="2">
  <i>24478</i>
  <i>24488</i>
  <uima.cas.FSArray>
```

Fig. 2 cTakes XML Example with Video Metadata. Example of XML source code from the cTakes result for a video related to blood cells



configuration uses the standard pipeline *AggregatePlain-textFastUMLSProcessor* to extract the SNOMED-CT terms.

To work with the UMLS library, we used a profile license.<sup>7</sup> Appendix 1 shows the configuration used to run the cTAKES execution.

Once the SNOMED-CT terms are extracted, we cross-match them with the terms from the Bio-ontology API to find synonymous MedlinePlus terms. These outputs allow us to obtain a web link from the *MP\_HEALTH\_TOPIC\_URL* MedlinePlus property, which is obtained via a Representational state transfer (REST) endpoint from the associated extracted term, which allows us to provide trusted recommendations to end users. For instance, the example terms *Blood* and *Stem Cell* both have corresponding Medline Plus links,<sup>8,9</sup>

Given that the number of SNOMED-CT vocabulary terms is larger than those on MedlinePlus, we anticipated that many results would not have matching terms. Although Bio-ontology offers an Annotator Web service that annotates user-provided text (e.g., journal abstracts) with relevant ontology concepts, this feature was not used for this work.

For practical reasons, we ignored isolated terms from SNOMED-CT that did not have a Medline Plus match. Although it is possible to select other ontologies to find a corresponding Medline Plus term, in this paper, we focus on results obtained only with these two ontologies.

**Datasets of videos and raters**

We assigned 26 health professionals (raters) to the three set of videos divided by topic (general medicine,

diabetes, or hypertension). We recruited these healthcare professionals directly either by email or other means, based on their familiarity with health topics and online health. After explaining to them the goals of the project and acquiring informed consent, the raters were asked to determine if the recommended links for a given video were relevant for the video topic. The exercise of rating the recommendations was not based on any personal information from the participants, but rather their expert opinion of a web tool (see Figs. 3 and 4). As such, this research does not involve human subjects (the study does not obtain information about living individuals).

Our dataset contained 53 videos, some of which had been utilized in our previous research [33]: a) 10 general medical videos (i.e., general health-related videos extracted from hospital YouTube channels), b) 22 videos about diabetes, and c) 21 videos about hypertension.

To rate the relevance of the videos and recommended links, we used Cohen's kappa to determine the level of agreement between two given raters. Kappa is defined as follows [42]:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{1}$$

where  $Pr(a)$  is the relative observed agreement and  $Pr(e)$  is the hypothetical chance of agreement. Therefore, this formula calculates the ratio of observed agreement

to hypothetical agreement by chance. If the raters are in complete agreement, then  $k = 1$ . A  $k$  coefficient greater than 0.80. indicates good agreement for a given recommendation.

Cohen's kappa was calculated using the *irr* package of the R application (version 3.3.1 on linux-gnu). The method in question is *kappa2(ratings, "unweighted")*. This function includes the vector of the rater values.

For each category of videos (general medical, diabetes, and hypertension), we selected a pair of reviewers with a high level of inter-rater agreement, based on Cohen's kappa, to have consistent rater agreement. The pair of raters had a Cohen's kappa inter-rater agreement of 0.626 for the general medical videos ( $z = 4.35$ ,  $p$ -value =  $1.33 \times 10^{-05}$ ), 0.582 for diabetes ( $z = 6.47$ ,  $p$ -value =  $9.9 \times 10^{-11}$ ), and 0.717 for hypertension ( $z = 7.7$ ,  $p$ -value =  $1.31 \times 10^{-14}$ ).

In the next step, we selected videos and links with an acceptable level of inter-rater agreement based on the Cohen's kappa values. Using the algorithm described in the previous section, we generated 510 recommended MedlinePlus links, but evaluated only the first five recommendations for each video, as our recommender system limits the number of recommendations. The final dataset contained 10 general medical videos with 48 recommended links, 22 diabetes videos with 102 recommended links, and 21 hypertension videos with 103 recommended links.



This approach allowed us to focus the evaluation on videos and links for which there was a homogenous agreement level among professionals. The rationale of this approach is relayed in our previous research, which highlighted the lack of consensus between professionals on certain types of health videos [43].

**Evaluation metrics of the recommendations**

We used two metrics to evaluate the relevance of the recommended links for a given video. These metrics, precision at  $k$  [44] and normalized discounted cumulative gain [45], are widely used to evaluate search algorithms in information retrieval and indicate the relevance of the “top” retrieved results. The importance of focusing on the top retrieved results is based on the web browsing behavior of users, as they tend to focus only on the top few item suggestions.

**Precision at  $k$**

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, in our case, this is the relevance of the links recommended for a given video. Precision is calculated as

$$\text{Precision} = \frac{|\text{Trusted Recommendations} \cap \text{Recovered Recommendations}|}{|\text{Recovered Recommendations}|} \tag{2}$$

The precision at  $k$  ( $P@k$ ) [46, 47] accounts for the order of the returned recommendations and is calculated as the fraction of the first  $k$  accepted links to all  $k$  links.

**Normalized Discounted Cumulative Gain**

The normalized discounted cumulative gain (nDCG) is another common information retrieval metric [45]. It is a measure of ranking quality, where  $DCG_k$  are highly relevant documents appearing lower in a search result and the ideal discounted cumulative gain ( $iDCG_k$ ) is the DCG of the vector with all links with an accepted value:

$$nDCG_k = \frac{DCG_k}{iDCG_k} \tag{3}$$

**Results**

To evaluate each recommendation, we considered two scenarios: a) robust and b) moderate. In the robust scenario, we consider as relevant only those link recommendations that are supported by both raters. In the moderate scenario, we consider a link to be relevant if at least one rater agreed with the recommendation. The moderate case is most appropriate when the risk of misinformation is low, while the robust scenario is the most appropriate when there is greater potential to spread misinformation.

In these scenarios,  $P@k$  and  $nDCG_k$  were calculated as follows. The relevance of the  $k$  first link recommendations is calculated as follows for each recommended link  $j$  ( $1 \leq j \leq k$ ):

- (a) If both raters approve link  $j$ , it is accepted (its value is 1, or *relevant*).
- (b) If both raters do not approve link  $j$ , it is rejected (its value is 0, or *irrelevant*).
- (c) In the case in which one rater approves link  $j$  and the other rejects it, in the robust scenario, link  $j$  is considered irrelevant (value 0), whereas in the moderate scenario, it is considered relevant (value 1).

The  $P@k$  results are shown in Table 1 and nDCG results are showed in Table 2. Overall, the performance of the recommender system was higher when giving recommendations for the general medicine and diabetes videos.

**Discussion**

The results show that it is feasible to recommend relevant links for health videos using a semantic-based recommender system. However, there are several concerns that deserve special attention. Although positive overall, recommendation performance varied across the different topics used in this study, which could be due several factors. For example, there might be fewer links related to diabetes than other topics (e.g., hypertension), thus limiting the potential items that can be recommended. Further, our semantic-based approach might also suffer from the semantic-gap between the layperson’s language and a medical thesaurus. Although work has been done to develop a Consumer Health Vocabulary, this has not been implemented in our approach; additionally, the semantic gap may differ across health topics [48].

In contrast, our approach of using semantics to identify relevant links allows the algorithms to find links that are related to synonyms and disambiguation. Still, this poses some additional challenges. For example, in a video titled *Juvenile Diabetes Research Foundation – Cure Video – Dalas*,<sup>10</sup> our algorithm extracted the term “shots,” which resulted in a recommendation for a link regarding the importance of vaccination (a topic of

**Table 1** Mean precision @ K recommended links

|                  | Mean Precision@k (robust case) |      |      | Mean Precision@k (moderate case) |      |      |
|------------------|--------------------------------|------|------|----------------------------------|------|------|
|                  | k=3                            | k=4  | k=5  | k=3                              | k=4  | k=5  |
| General Medicine | 0.77                           | 0.65 | 0.5  | 0.87                             | 0.80 | 0.70 |
| Diabetes         | 0.71                           | 0.71 | 0.68 | 0.89                             | 0.85 | 0.81 |
| HTN              | 0.48                           | 0.45 | 0.39 | 0.62                             | 0.57 | 0.53 |

The evaluation based on nDCG (see Table 2) shows similar patterns, and lower performance when recommending links for hypertension videos. As expected, the relevance of the links decreased with an increase in the number of recommended links for a given video ( $k=5$ )

**Table 2** Mean nDCG for K recommended links

|                  | Mean nDCG <sub>k</sub><br>(robust case) |       |       | Mean nDCG <sub>k</sub><br>(moderate case) |       |       |
|------------------|---|-------|-------|---|-------|-------|
|                  | k = 3                                   | k = 4 | k = 5 | k = 3                                     | k = 4 | k = 5 |
| General Medicine | 0.78                                    | 0.7   | 0.5   | 0.88                                      | 0.83  | 0.75  |
| Diabetes         | 0.73                                    | 0.74  | 0.72  | 0.90                                      | 0.87  | 0.85  |
| HTN              | 0.51                                    | 0.49  | 0.46  | 0.65                                      | 0.61  | 0.58  |

relative importance in diabetes). One advantage of relying on medical terms is that our algorithm has an enhanced capability to reduce the number of links that have no relation to the video content, which is an important limitation of previous studies, where such terms could not be avoided [33, 34].

Recommender systems can play a major role, not only in education, but also in supporting behavioral changes for a wide range of health conditions [49–51], including smoking cessation [51]. In such cases, the recommendations are not only chosen with regard to content, but also with respect to timing, and consider different psychological health factors (aka user context) [52]. Our work does not address context-awareness regarding the time and place of the recommendations. However, by providing trustworthy recommendations for websites when a user is watching a video, we can support complex health information seeking [53, 54].

The applications of recommender systems in the health domain are still emerging. Therefore, we lack common evaluation methods that can allow us to compare work across separate studies in this topic [29, 55]. There are examples in the literature of recommender systems in the health domain that, for example, provide recommendations based on a personal health record [56]. In our case, we deal with a very different type of content-based recommendation, as we are not recommending content for a given user but rather for a given health educational item.

Our work is aligned with previous studies in which health information is enriched with additional content [56]. There is still quite a substantial knowledge gap on how people search for online health information, and, even more importantly, on how that affects the health behaviors of the information seeker [57]. Our recommender system approach does not aim to provide recommendations personalized for a user, but rather to provide further reliable information for users watching a health video. This content-based recommendation approach is crucial for supporting the current patterns of health consumers looking for multiple sources when searching for health information online [58].

Most previous studies of health recommender systems do not address their impact on health outcomes; in

contrast, we do so using information retrieval accuracy metrics. This approach has the potential to create risks for health consumers, which is one of our motivations for using health professionals in this evaluation. Ekstrand et al. recently reviewed potential ways in which health recommender systems can do harm and the ways to minimize potential harm [59]. Giving wrong or potentially misleading health information can be a cause for serious concern; for example, recently, the FDA forced the company 23andMe to remove and edit personalized health information regarding genetic health risks [60]. Further, health information can be used for unhealthy purposes (e.g., the abuse of diuretics for weight loss is common in people with eating disorders).

### Limitations

Our study relies on the ratings of hundreds of recommended links for given videos. However, these ratings were given by healthcare professionals and not health consumers. As explained in our previous work, professionals and consumers often disagree on the relevance of health content [43]. Experiments with health consumers will be required to further evaluate recommendation quality.

Note that our study only investigates the feasibility of this approach. Consequently, extrapolating the results to larger studies is necessary. Ideally, further studies will consider more users (and not necessarily healthcare professionals). In addition, our rating approach was rather simplistic, considering the multiple quality dimensions of health videos [35]. Further, the ideal evaluation should take place in a real information seeking scenario and not a simulated one because many factors affect information seeking by health consumers, including stress or literacy levels [53]. The patient perspective was not explored in this study because we consider it to be more ethically appropriate to first study the feasibility of an approach with health experts. Patients' perspectives and acceptance can also vary substantially across age, health literacy levels, and other factors. Future research will need to explore the application of our method in a patient portal with additional content and users.

Another limitation of our study is that our video dataset is not generalizable. We selected several topics of high importance (diabetes and hypertension), but we cannot extrapolate that our approach will work with other health topics. A major challenge to generalizing semantic-based approaches such as ours is the gap between medical and consumer health vocabularies [61]. Because we use content generated by health organizations (not individuals) and a medical ontology, we might expect more difficulties when recommending links to consumer-generated content.



## Conclusions

This study demonstrated that a semantic-based recommender algorithm can provide relevant education health websites as further reading for a given health video. The relevance of websites recommended by our system decreased as we provided more recommendations, but HealthRecSys still performed well with up to five recommended links per video. Because user browsing behavior is often limited to a few items, this does not pose a serious limitation. Conversely, our approach can reduce the burden of health consumers when searching for reliable additional health educational content. Further, the speed of navigation to a reliable source, as identified by Strauss, is an important factor in information seeking [62].

Future improvements to recommender systems will incorporate more semantic analytics and perhaps be able to determine the patient's context (i.e., mood) to make better recommendations. It will be possible to use this algorithm to recommend content and videos to counterbalance misinformation, find information on controversial topics, and filter out videos with little scientific acceptance. For instance, a video that promotes steroid consumption could recommend information alerting the individual to their potential negative effects.

## Endnotes

<sup>1</sup>Bio-ontology API endpoint documentation <http://data.bioontology.org/documentation>

<sup>2</sup>Bio-ontology website <http://www.bioontology.org/>

<sup>3</sup>SNOMED-CT website <http://www.ihtsdo.org/snomed-ct>

<sup>4</sup><http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html>

<sup>5</sup>cTAKES website <http://ctakes.apache.org/>

<sup>6</sup>UMLS website <https://www.nlm.nih.gov/research/umls/>

<sup>7</sup>UMLS web license profile <https://uts.nlm.nih.gov/uts.html#profile>

<sup>8</sup>Blood: <https://www.nlm.nih.gov/medlineplus/blood.html>

<sup>9</sup>Cell: <https://www.nlm.nih.gov/medlineplus/stemcells.html>

<sup>10</sup>Juvenile Diabetes Research Foundation <https://www.youtube.com/watch?v=i7ft-6vR-1c>

## Additional file

**Additional file 1:** HealthRecSys Study Overview. Video describing the HealthRecSys algorithm and the results of the study.

## Abbreviations

API: Application-programming interface; NLP: Natural language processing; UMLS: Unified Medical Language System

## Acknowledgements

We would like to acknowledge the efforts of the video reviewers who completed the evaluations that enabled us to complete this study.

## Funding

This research was co-funded by a HealthTrust research grant from the University of Tromsø (Norway). This study was also co-funded by the SmokeFreeBrain Project of the European Union's Horizon 2020 Research and Innovation Programme under Grant No. 681120. The funding sources did not have any role in the execution of the study.

## Availability of data and materials

The dataset supporting the conclusions of this article is included within the article.

## Authors' contributions

CLS led the study and consequently was involved in all research and manuscript preparation. LFL, AC, and JSR reviewed the manuscript and advised CLS on the algorithm used in the study. CR participated in the study design, discussions, and content analysis, and reviewed the manuscript. All the authors read and approved the final version of the manuscript.

## Authors' information

LFL performed part of this work at the Northern Research Institute (Norut) in Tromsø, Norway.

## Competing interests

The authors declare that they have no competing interests.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

Not applicable. The experts that participated in the study did not provide any personal information, but rather evaluated a tool; consequently, their participation is not considered to be human subject research. The research protocol of this study was evaluated by the PhD Research Committee of the University of Seville, Spain, and all collaborators provided informed written consent.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Department of Architecture and Computer Technology Universidad de Sevilla, Seville, Spain. <sup>2</sup>eHealth Researcher, Toronto, Canada. <sup>3</sup>Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar Foundation, PO Box 5825, Doha, Qatar.

Received: 11 November 2016 Accepted: 24 March 2017

## References

- Griffis HM, Kilaru AS, Werner RM, Asch DA, Hershey JC, Hill S, Ha YP, Sellers A, Mahoney K, Merchant RM. Use of social media across US hospitals: descriptive analysis of adoption and utilization. *J Med Internet Res*. 2014; 16(11):e264. doi:10.2196/jmir.3758.
- Ziebland S, Wyke S. Health and illness in a connected world: how might sharing experiences on the internet affect people's health? *Milbank Q*. 2012; 90:219–49.
- Jamal A, Khan SA, AlHumud A, Al-Duhayyim A, Alrashed M, Bin Shabr F, et al. Association of online health information-seeking behavior and self-care activities among type 2 diabetic patients in Saudi Arabia. *J Med Internet Res*. 2015;17:e196.
- Lewis D. What is Web 2.0? Crossroads. 2006;13:3.
- Zeng X, Parmanto B. Web content accessibility of consumer health information web sites for people with disabilities: a cross sectional evaluation. *J Med Internet Res*. 2004;6:e19.
- Cline RJW, Haynes KM. Consumer health information seeking on the Internet: the state of the art. *Health Educ Res*. 2001;16:671–92.
- Fernandez-Luque L, Karlsen R, Vognlid LK. Challenges and opportunities of using recommender systems for personalized health education. In *Stud Health Technol Inform*. 2009;150:903–7. doi:10.3233/978-1-60750-044-5-903.

8. Ponce V, Deschamps JP, Giroux LP, Salehi F, Abdulrazak B, QueFaire: Context-aware in-person social activity recommendation system for active aging. In: Geissbühler A, Demongeot J, Mokhtari M, Abdulrazak B, Aloulou H, editors. *Inclusive Smart Cities and e-Health*. Springer International Publishing; 2015. p. 64–75.
9. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE T Knowl Data Eng*. 2005;17:734–49.
10. Sanchez-Bocanegra CL, Sanchez-Laguna F, Sevillano JL. Introduction on health recommender systems. *Methods Mol Biol* (Clifton NJ). 2015;1246: 131–46.
11. Xavier A. Past, Present, and Future of Recommender System. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ACM 2016. <https://doi.org/10.1145/2959100.2959144>.
12. Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: *UAI'98 Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, 1998. p. 43–52.
13. Elahi M, Ricci F, Rubens N. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*. 2016;20:29–50.
14. Gu Y, Zhao B, Hardtke D, Sun Y. Learning Global Term Weights for Content-based Recommender Systems. In: *International World Wide Web Conference Committee (IW3C2)*. Canada: ACM; 2016.
15. Zhang H-R, Min F. Three-way recommender systems based on random forests. *Knowl-Based Syst*. 2016;91:275–86.
16. Li Y, Zheng Y, Kang J, Bao H. Designing a Learning Recommender System by Incorporating Resource Association Analysis and Social Interaction Computing. In: Li Y, Chang M, Kravcik M, Popescu E, Huang R, Kinshuk, Chen N-S, editors. *Springer State-of-the-Art and Future Directions of Smart Learning*. Singapore: Springer; 2015. p. 137–43.
17. Campos L, Fernández-Luna J, Huete J, Rueda-Morales M. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. *Int J Approx Reason*. 2010;51:785–99.
18. Pham T-N, Vuong T-H, Thai T-H, Tran M-V, Ha Q-T. Sentiment Analysis and User Similarity for Social Recommender System: An Experimental Study. *Information Science and Applications (ICISA)*. 2016;376:1147–56.
19. Aggarwal C. Social and Trust-Centric Recommender Systems. *Recommender Systems*. 2016. p. 345–84.
20. Doerfel S, Jäschke R, Stumme G. The Role of Cores in Recommender Benchmarking for Social Bookmarking Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2016;7:1–33. <https://doi.org/10.1145/2700485>.
21. Jiang W, Wang G, Bhuiyan MZ, Wu J. Understanding Graph-Based Trust Evaluation in Online Social Networks. *ACM Computing Surveys (CSUR)*. 2016;49:1–35.
22. Felfernig A, Friedrich G, Jannach D, Zanker M. Constraint-based recommender systems. *Recommender Systems Handbook 2016*;161–90. doi: 10.1007/978-1-4899-7637-6\_5.
23. Bridge D, Goker M, McGinty L, Smyth B. Case-based recommender systems. *Knowl Eng Rev*. 2005;20(3):315–20.
24. Garrido A, Morales L, Serina I. On the use of case-based planning for e-learning personalization. *Expert Systems with Applications*. 2016;60:1–15.
25. Hernandez A, Bobadilla J, Ortega F. A non-negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowl-Based Syst*. 2016;97:188–202.
26. Ioannidis E, Weinsberg E, Taft NA, Joye M, Nikolaenko V, inventors; Thomson Licensing, assignee. A method and system for privacy-preserving recommendation to rating contributing users based on matrix factorization. United States patent US 20,160,012,238. 24 Jan 2016.
27. Espin V, Hurtado MV, Noguera M. Nutrition for elder care: A nutritional semantic recommender system for the elderly. *Expert Syst*. 2016;33:201–10.
28. Borbolla D, Del Fiol G, Taliercio V, Otero C, Campos F, Martinez M, Luna D, Quiros F. Integrating personalized health information from MedlinePlus in a patient portal. *Stud Health Technol Inform*. 2014;205:348–52.
29. Valdez AC, Zieffe M, Verbert K, Felfernig A, Holzinger A. Recommender systems for health informatics: state-of-the-art and future perspectives. *Machine Learning for Health Informatics*. Springer International Publishing, 2016, p. 391–414.
30. Berendt B, Hotho A, Stumme G. Towards semantic Web mining. In: Horrocks I, Hendler J, editors. *The Semantic Web—ISWC 2002*. Heidelberg: Springer; 2002. p. 264–78.
31. Berners-Lee T, Hendler J. Publishing on the semantic web. *Nature*. 2001; 410(6832):1023–4.
32. Yu HQ, Zhao X, Deng Z, Dong F. Ontology driven personal health knowledge discovery. In: Uden L, Heričko M, Ting I-H, editors. *Knowledge Management in Organizations*. Springer International Publishing; 2015. p. 649–63.
33. Rivero-Rodríguez A, Konstantinidis ST, Sanchez-Bocanegra CL, Fernandez-Luque L. A health information recommender system: enriching YouTube health videos with MedlinePlus information by the use of SNOMEDCT terms. In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, Porto, Portugal, June. IEEE. 2013. p. 257–61.
34. Sánchez-Bocanegra CL, Rivero-Rodríguez A, Fernández-Luque L, Sevillano JL. Diavideos: A diabetes health video portal. *Stud Health Technol Inform*. 2013;2(1):e6.
35. Gabarron E, Fernandez-Luque L, Armayones M, Lau AY. Identifying measures used for assessing quality of YouTube videos with patient health information: a review of current literature. *Interact J Med Res*. 2013;2(1):e6. doi:10.2196/ijmr.2465.
36. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544–51. doi: 10.1136/amiajnl-2011-000464.
37. Alexander C, Chris F, Shalom L. *The Handbook of Computational Linguistics and Natural Language Processing*. USA: Wiley-Blackwell; 2010. ISBN 978-1-4051-5581-6.
38. Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: A case study of warfarin. *J Am Med Inform Assoc*. 2011;18:387–91.
39. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13:395–405.
40. Uzuner O, Stubbs A. Practical applications for natural language processing in clinical research: The 2014 I2b2/UTHealth shared tasks. *J Biomed Inform*. 2015;58:51–5.
41. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: a prospective analysis on statewide electronic medical records. *Int J Med Inf*. 2015;84:1039–47.
42. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257–68.
43. Fernandez-Luque L, Karlson R, Melton GB. HealthTrust: a social network approach for retrieving online health videos. *J Med Internet Res*. 2012;14(1): e22. doi:10.2196/jmir.1985.
44. Sujatha P, Dhavachelvan P. Precision at K in multilingual information retrieval. *Int J Comput Appl*. 2011;24:40–3.
45. Wang Y, Wang L, Li Y, He D, Chen W, Liu TY. A theoretical analysis of normalized discounted cumulative gain (NDCG) ranking measures. In: *Proc 26<sup>th</sup> Annu Conf Learning Theory*. 2013. p. 1–30.
46. Kobayashi M, Takeda K. Information retrieval on the Web. *ACM Comput Surv*. 2000;32(2):144–73.
47. Chapelle O, Wu M. Gradient descent optimization of smoothed information retrieval metrics. *Inf Retr*. 2009;13:216–35.
48. Zeng Q, Tse T, Divita G, Keselman A, Crowell J, Browne AC. Term identification methods for consumer health vocabulary development. *J Med Internet Res*. 2007;9(1):e4.
49. Wiesner M, Pfeifer D. Health recommender systems: Concepts, requirements, technical basics and challenges. *Int J Environ Res Publ Health*. 2014;11(3):2580–607.
50. Sadasivam RS, Cutrona SL, Kinney RL, Marlin BM, Mazor KM, Lemon SC, et al. Collective-Intelligence recommender systems: Advancing computer tailoring for health behavior change into the 21st century. *J Med Internet Res*. 2016; 18(3):e42.
51. Hors-Fraile S, Civit A, Nuñez Benjumea FJ, Carrasco Hernández L, Ortega Ruiz F, Fernandez-Luque L. Coupling neuroscience smoking cessation interventions with social media and mobile devices. *Front Hum Neurosci*. Conference Abstract: SAN2016 Meeting. 2016. doi:10.3389/conf.fnhum.2016.220.00025.
52. Lin Y, Jessurun AJ, de Vries B, Timmermans HJP. Motivate: towards context-aware recommendation mobile system for healthy living. In: *Proc IEEE 5<sup>th</sup> Int Conf Pervasive Computing Technologies for Healthcare*. 2011. p. 250–3.
53. Lau AYS, Coiera EV. Can cognitive biases during consumer health information searches be reduced to improve decision making? *J Am Med Inform Assoc*. 2009;16(1):54–65.



54. Sharit J, Taha J, Berkowsky RW, Czaja SJ. Seeking and Resolving Complex Online Health Information Age Differences in the Role of Cognitive Abilities. *Proc Human Factors and Ergonomics Soc Annu Meeting*. 2016;60:1.
55. Elsweller D, Ludwig B, Said A, Schaefer H, Trattner C. Engendering Health with Recommender Systems. In: *Proc 10<sup>th</sup> ACM Conf Recommender Systems*. 2016. p. 409–10.
56. Wiesner M, Pfeifer D. Adapting recommender systems to the requirements of personal health record systems. In: Veinot T, editor. *Proc 1<sup>st</sup> ACM Int Health Inform Symp (IHIT'10)*. USA: ACM; 2010. p. 410–4.
57. Anker AE, Reinhart AM, Feeley TH. Health information seeking: a review of measures and methods. *Patient Educ Couns*. 2011;82(3):346–54.
58. Longo DR, Schubert SL, Wright BA, LeMaster J, Williams CD, Clore JN. Health information seeking, receipt, and use in diabetes self-management. *Ann Family Med*. 2010;8(4):334–40.
59. Ekstrand JD, Ekstrand MD. First do no harm: Considering and minimizing harm in recommender systems designed for engendering health. *Engendering Health Workshop RecSys 2016 Conf*. 2016. <https://md.ekstrandom.net/research/pubs/do-no-harm/FirstDoNoHarm.pdf> Accessed 2 Apr 2017.
60. Kill A. The direct-to-consumer genetics debate. *Lancet Oncol*. 2016;17(7):e265.
61. Zeng QT, Tony T, Guy D, Alla K, Jon C, Browne AC, Sergey G, Long N. Term Identification Methods for Consumer Health Vocabulary Development. *J Med Internet Res*. 2007;9(1):e4.
62. Straus S, Haynes RB. Managing evidence-based knowledge: the need for reliable, relevant and readable resources. *CMAJ Can Med Assoc J*. 2009;180: 942–5.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)





# Bibliografía

- [Adomavicius and Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.
- [Agarwal et al., 2016] Agarwal, V., Zhang, L., Zhu, J., Fang, S., Cheng, T., Hong, C., and Shah, H. N. (2016). Impact of predicting health care utilization via web search behavior: A data-driven analysis. *J Med Internet Res*, 18(9):e251.
- [Ahram et al., 2015] Ahram, T., Karwowski, W., Schmorrow, D., Razza, B., and Paschoarelli, L. C. (2015). Evaluation of usability aspects of consumer products in online sales. *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015)*, 3:6237 – 6244.
- [Akerkar et al., 2012] Akerkar, R., Bădică, C., and Burdescu, D. D. (2012). Desiderata for research in web intelligence, mining and semantics. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 0. ACM.
- [Andrieu et al., 2003] Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43.
- [Banos et al., 2015] Banos, O., Amin, M. B., Khan, W. A., Afzel, M., Ahmad, M., Ali, M., Ali, T., Ali, R., Bilal, M., Han, M., Hussain, J., Hussain, M., Hussain, S., Hur, T. H., Bang, J. H., Huynh-The, T., Idris, M., Kang, D. W., Park, S. B., Siddiqui, H., Vui, L.-B., Fahim, M., Khattak,

- A. M., Kang, B. H., and Lee, S. (2015). *An Innovative Platform for Person-Centric Health and Wellness Support*, pages 131–140. Springer International Publishing, Cham.
- [Barbieri and Manco, 2011] Barbieri, N. and Manco, G. (2011). *An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering*, pages 172–187. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Basagoiti, 2012] Basagoiti, I. (2012). *Alfabetización en salud. De la información a la acción [pdf]*. Valencia.
- [Bedi et al., 2007] Bedi, P., Kaur, H., and Marwaha, S. (2007). Trust based recommender system for the semantic web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*, pages 2677–2682, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Benesty et al., 2009] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). *Pearson Correlation Coefficient*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Benslimane et al., 2008] Benslimane, D., Dustdar, S., and Sheth, A. P. (2008). Services mashups: The new generation of web applications. *IEEE Internet Computing (INTERNET)*, 12(5):13–15.
- [Benson, 2009] Benson, T. (2009). *Principles of Health Interoperability HL7 and SNOMED (Health Informatics)*. Springer, first edition.
- [Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.
- [Bridge et al., 2005] Bridge, D., Göker, M. H., McGinty, L., and Smyth, B. (2005). Case-based recommender systems. *Knowl. Eng. Rev.*, 20(3):315–320.
- [Broder and Ciccolo, 2004] Broder, A. Z. and Ciccolo, A. C. (2004). Towards the next generation of enterprise search technology. *IBM Systems Journal*, 43(3):451–454.

- [Burgess et al., 2013] Burgess, J., Green, J., Jenkins, H., and Hartley, J. (2013). *YouTube: Online Video and Participatory Culture*. Digital Media and Society. Wiley.
- [Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.
- [Burke et al., 2006] Burke, R., Mobasher, B., Williams, C., and Bhaumik, R. (2006). Classification features for attack detection in collaborative recommender systems. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–547. ACM.
- [Büttcher et al., 2016] Büttcher, S., Clarke, C. L., and Cormack, G. V. (2016). *Information retrieval: Implementing and evaluating search engines*. Mit Press.
- [Capelle et al., 2012] Capelle, M., Frasincar, F., Moerland, M., and Hogenboom, F. (2012). Semantics-based news recommendation. In *Proceedings of the 2nd international conference on web intelligence, mining and semantics*, page 27. ACM.
- [Carbonell, 1994] Carbonell, J. (1994). El procesamiento del lenguaje natural, tecnología en transición. In *Actas del Congreso de la Lengua Española: Sevilla, 7 al 10 octubre, 1992*, pages 247–250. Instituto Cervantes.
- [Carletta, 1996] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.
- [Carroll et al., 2004] Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. (2004). Jena: implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, WWW Alt. 2004, pages 74–83, New York, NY, USA. ACM.
- [Chapelle and Wu, 2009] Chapelle, O. and Wu, M. (2009). Gradient descent optimization of smoothed information retrieval metrics. *Inf Retrieval*, 13(3):216–235.

- [Cheng et al., 2008] Cheng, X., Dale, C., and Liu, J. (2008). Statistics and social network of youtube videos. In *16th International Workshop on Quality of Service. IWQoS 2008.*, pages 229–238. IEEE.
- [Cheng et al., 2012] Cheng, X., Liu, J., Wang, H., and Wang, C. (2012). Coordinate live streaming and storage sharing for social media content distribution. *IEEE Transactions on Multimedia*, 14(6):1558–1565.
- [Chou et al., 2011] Chou, W.-Y. S., Hunt, Y., Folkers, A., and Augustson, E. (2011). Cancer Survivorship in the Age of YouTube and Social Media: A Narrative Analysis. *Journal of Medical Internet Research*, 13(1):e7.
- [Choudhury et al., 2014] Choudhury, M. D., Morris, M. R., and White, R. W. (2014). Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems (CHI 2014), pages 1365–1376, New York, NY, USA. ACM.
- [Chute et al., 2010] Chute, C. G., Beck, S. A., Fisk, T. B., and Mohr, D. N. (2010). The enterprise data trust at mayo clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association*, 17(2):131–135.
- [Cline and Haynes, 2001] Cline, R. J. W. and Haynes, K. M. (2001). Consumer health information seeking on the Internet: the state of the art. *Health Educ. Res.*, 16(6):671–692.
- [Cushman et al., 2010] Cushman, R., Froomkin, A. M., Cava, A., Abril, P., and Goodman, K. W. (2010). Ethical, legal and social issues for personal health records and applications. *Journal of Biomedical Informatics*, 43(5, Supplement):S51 – S55. Project HealthDesign.
- [Daniulaityte et al., 2016] Daniulaityte, R., Chen, L., Lamy, R. F., Carlson, G. R., Thirunarayan, K., and Sheth, A. (2016). When ‘bad’ is ‘good’: Identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health Surveill*, 2(2):e162.

- [Davidson et al., 2010] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM.
- [Davoodi et al., 2013] Davoodi, E., Kianmehr, K., and Afsharchi, M. (2013). A semantic social network-based expert recommender system. *Applied intelligence*, 39(1):1–13.
- [Ding et al., 2011] Ding, B., Winslett, M., Han, J., and Li, Z. (2011). Differentially private data cubes: optimizing noise sources and consistency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 217–228. ACM.
- [Donsbach et al., 2010] Donsbach, A., Gregov, A., Few, J. D., Lehman, J. D., and Brownell, J. T. (2010). Recommendations based on item tagging activities of users. US Patent 7,668,821.
- [Dumais, 1991] Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.
- [Ekstrand, 2014] Ekstrand, M. D. (2014). *Towards recommender engineering tools and experiments for identifying recommender differences*. PhD thesis, UNIVERSITY OF MINNESOTA.
- [Erekhinskaya et al., 2016] Erekhinskaya, T., Balakrishna, M., Tatu, M., and Moldovan, D. (2016). *Personalized Medical Reading Recommendation: Deep Semantic Approach*, pages 89–97. Springer International Publishing, Cham.
- [Eysenbach, 2008] Eysenbach, G. (2008). Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res*, 10(3):e22.
- [Fernandez-Luque et al., 2011] Fernandez-Luque, L., Karlsen, R., and Bonander, J. (2011). Review of extracting information from the social web for health personalization. *J Med Internet Res*, 13(1):e15.

- [Fernandez-Luque et al., 2012] Fernandez-Luque, L., Karlsen, R., and Melton, G. B. (2012). Healthtrust: A social network approach for retrieving online health videos. *Journal of Medical Internet Research*, 14(1):e22.00010.
- [Fernandez-Luque et al., 2009] Fernandez-Luque, L., Karlsen, R., and Vogt, L. K. (2009). Challenges and opportunities of using recommender systems for personalized health education. In *Medical Informatics (MIE)*, pages 903–907.
- [Fernandez-Luque L., 2014] Fernandez-Luque L., Sánchez-Bocanegra C., R. J. K. R. (2014). A semantic web health recommender system: Enriching youtube health videos. In *7th World Congress on Social Media, Mobile Apps, Web 2.0 in Health and Medicine - Medicine 2.0*.
- [Fernández-Tobías, 2017] Fernández-Tobías, I. (2017). Matrix factorization models for cross-domain recommendation: Addressing the cold start in collaborative filtering.
- [Florian et al., 2003] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, pages 168–171, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Fox et al., 2012] Fox, R., Cooley, J., McGrath, M., and Hauswirth, M. (2012). Mobile health apps - from singular to collaborative. *Stud Health Technol Inform*, 177:158–163.
- [Fox and Jones, 2011] Fox, S. and Jones, S. (2011). The social life of health information. <http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information>.
- [Fung and Yee, 1998] Fung, P. and Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98*, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.



- [Gao et al., 2015] Gao, H., Bowles, E. J. A., Carrell, D., and Buist, D. S. (2015). Using natural language processing to extract mammographic findings. *Journal of Biomedical Informatics*, 54:77 – 84.
- [Gavalas et al., 2014] Gavalas, D., Konstantopoulos, C., Mastakas, K., and Pantziou, G. (2014). Review: Mobile recommender systems in tourism. *J. Netw. Comput. Appl.*, 39:319–333.
- [Givon, 2011] Givon, S. (2011). Predicting and using social tags to improve the accuracy and transparency of recommender systems.
- [Guan et al., 2009] Guan, Z., Bu, J., Mei, Q., Chen, C., and Wang, C. (2009). Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 540–547. ACM.
- [Hameed, 2003] Hameed, K. (2003). The application of mobile computing and technology to health care services. *Telemat. Inf.*, 20(2):99–106.
- [Harrington, 2012] Harrington, P. (2012). *Machine learning in action*, volume 5. Manning Greenwich, CT.
- [Hipp et al., 2000] Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining - a General Survey and Comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64.
- [Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.
- [Huang et al., 2004a] Huang, Z., Chung, W., and Chen, H. (2004a). A graph model for e-commerce recommender systems. *J. Am. Soc. Inf. Sci. Technol.*, 55(3):259–274.
- [Huang et al., 2004b] Huang, Z., Chung, W., and Chen, H. (2004b). A graph model for e-commerce recommender systems. *Journal of the Association for Information Science and Technology*, 55(3):259–274.

- [Hurley, 2011] Hurley, N. J. (2011). Robustness of recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 9–10. ACM.
- [Iaquinta et al., 2008] Iaquinta, L., d. Gemmis, M., Lops, P., Semeraro, G., Filannino, M., and Molino, P. (2008). Introducing serendipity in a content-based recommender system. In *Eighth International Conference on Hybrid Intelligent Systems. HIS 2008*, pages 168–173.
- [Jannach et al., 2010] Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition.
- [Jäschke et al., 2007] Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007). *Tag Recommendations in Folksonomies*, pages 506–514. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Jindal and Liu, 2008] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM.
- [Karatzoglou et al., 2016] Karatzoglou, A., Hidasi, B., Tikk, D., Sarshalom, O., Roitman, H., Shapira, B., and Rokach, L. (2016). Recsys’16 workshop on deep learning for recommender systems (dlrs). In *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys ’16*, pages 415–416, New York, NY, USA. ACM.
- [Kietzmann et al., 2011] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251.
- [Kimmel et al., 2005] Kimmel, Z., Greenes, R., and Liederman, E. (2005). Personal health records. *J Med Pract Manage*, 21(3):147–152.
- [Kobayashi and Takeda, 2000] Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173.
- [Konstantinidis et al., 2013] Konstantinidis, S., Fernandez-Luque, L., Bamidis, P., and Karlsen, R. (2013). The role of taxonomies in social media

- and the semantic web for health education. A study of SNOMED CT terms in YouTube health video tags. *Methods Inf Med*, 52(2):168–179.
- [Koren et al., 2009] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37.
- [Kwon, 2000] Kwon, G. (2000). *Rewrite Rules and Operational Semantics for Model Checking UML Statecharts*, pages 528–540. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Landauer et al., 1998] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.
- [Lewis et al., 1996] Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R. (1996). Training algorithms for linear text classifiers. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96*, pages 298–306, New York, NY, USA. ACM.
- [Li et al., 2012] Li, X., Su, X., and Wang, M. (2012). Social network-based recommendation: a graph random walk kernel approach. In *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries*, pages 409–410. ACM.
- [Linden et al., 2003] Linden, G., Smith, B., and York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1):76–80.
- [Liu et al., 2013] Liu, H., Waghlikar, K., Jonnalagadda, S., and Sohn, S. (2013). Integrated ctakes for concept mention detection and normalization. In *Working Notes for CLEF 2013 Conference*, volume 1179. CEUR-WS.
- [Lops et al., 2011] Lops, P., Gemmis, M. d., and Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 73–105. Springer US.

- [Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.
- [Marin-Torres Viviana et al., 2013] Marin-Torres Viviana, Valverde Aliaga Justo, Sánchez Miró Ignacio, Sáenz del Castillo Vicente María Isabel, Polentinos-Castro Elena, and Garrido Barral Araceli (2013). Internet como fuente de información sobre salud en pacientes de atención primaria y su influencia en la relación médico-paciente. *Atención Primaria*, 45(1):46–53.
- [Martínez-Torres, 2015] Martínez-Torres, M. R. (2015). Content analysis of open innovation communities using latent semantic indexing. *Technology Analysis & Strategic Management*, 27(7):859–875.
- [McKenzie et al., 2016] McKenzie, J. F., Neiger, B. L., and Thackeray, R. (2016). *Planning, implementing & evaluating health promotion programs: A primer*. Pearson.
- [McMullan, 2006] McMullan, M. (2006). Patients using the internet to obtain health information: How this affects the patient–health professional relationship. *Patient Education and Counseling*, 63(1–2):24 – 28.
- [Melville and Sindhvani, 2010] Melville, P. and Sindhvani, V. (2010). *Recommender Systems*, pages 829–838. Springer US, Boston, MA.
- [Mirza, 2001] Mirza, B. J. (2001). *Jumping connections: A graph-theoretic model for recommender systems*. PhD thesis, Virginia Tech.
- [Myers and Sirois, 2014] Myers, L. and Sirois, M. J. (2014). *Spearman Correlation Coefficients, Differences between*. John Wiley & Sons, Ltd.
- [Nazi, 2013] Nazi, M. K. (2013). The personal health record paradox: Health care professionals’ perspectives and the information ecology of personal health record systems in organizational and clinical settings. *J Med Internet Res*, 15(4):e70.
- [Nguyen and Bai, 2011] Nguyen, H. V. and Bai, L. (2011). *Cosine Similarity Metric Learning for Face Verification*, pages 709–720. Springer Berlin Heidelberg, Berlin, Heidelberg.

- [Noveck, 2009] Noveck, B. S. (2009). Wiki government: How technology can make government better, democracy stronger, and citizens more powerful. <http://ictlogy.net/bibliography/reports/projects.php?idp=1737>.
- [O'Donovan and Smyth, 2006] O'Donovan, J. and Smyth, B. (2006). Is trust robust?: an analysis of trust-based recommendation. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 101–108. ACM.
- [Olney et al., 2007] Olney, C. A., Warner, D. G., Reyna, G., Wood, F. B., and Siegel, E. R. (2007). MedlinePlus and the challenge of low health literacy: findings from the Colonias project. *Journal of the Medical Library Association : JMLA*, 95(1):31–39.
- [Oxman and Guyatt, 1993] Oxman, A. D. and Guyatt, G. H. (1993). The science of reviewing research. *Annals of the New York Academy of Sciences*, 703(1):125–134.
- [Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.
- [Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02*, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Park et al., 2015] Park, R. C., Jung, H., Shin, D.-K., Kim, G.-J., and Yoon, K.-H. (2015). M2m-based smart health service for human ui/ux using motion recognition. *Cluster Computing*, 18(1):221–232.
- [Pattaraintakorn and Cercone, 2008] Pattaraintakorn, P. and Cercone, N. (2008). A foundation of rough sets theoretical and computational hybrid intelligent system for survival analysis. *Computers & Mathematics with Applications*, 56(7):1699–1708.
- [Pazzani and Billsus, 2007] Pazzani, M. J. and Billsus, D. (2007). *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg.

- [Peis et al., 2008] Peis, E., del Castillo, J. M. M., and Delgado-López, J. A. (2008). Semantic recommender systems. analysis of the state of the topic. *Hipertext.net*, 6.
- [Peng et al., 2007] Peng, F., Ahmed, N., Li, X., and Lu, Y. (2007). Context sensitive stemming for web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 639–646, New York, NY, USA. ACM.
- [Rector et al., 2009] Rector, A. L., Qamar, R., and Marley, T. (2009). Binding ontologies and coding systems to electronic health records and messages. *Appl. Ontol.*, 4(1):51–69.
- [Ricci, 2010] Ricci, F. (2010). Mobile recommender systems. *Information Technology & Tourism*, 12(3):205–231.
- [Rivero-Rodriguez et al., 2013] Rivero-Rodriguez, A., Konstantinidis, S. T., Sánchez-Bocanegra, C. L., and Fernández-Luque, L. (2013). A health information recommender system: Enriching youtube health videos with medline plus information by the use of snomedct terms. In Rodrigues, P. P., Pechenizkiy, M., Gama, J., Cruz-Correia, R., Liu, J., Traina, A. J. M., Lucas, P. J. F., and Soda, P., editors, *IEEE International Symposium on Computer-Based Medical Systems, (CBMS 2013)*, pages 257–261. IEEE Computer Society.
- [Roitman et al., 2010a] Roitman, H., Messika, Y., Tsimmerman, Y., and Maman, Y. (2010a). Increasing patient safety using explanation-driven personalized content recommendation. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 430–434, New York, NY, USA. ACM.
- [Roitman et al., 2010b] Roitman, H., Messika, Y., Tsimmerman, Y., and Maman, Y. (2010b). Increasing patient safety using explanation-driven personalized content recommendation. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 430–434. ACM.
- [Salcedo, 2011] Salcedo, V. T. (2011). *El ePaciente y las redes sociales*. ITACA-TSB.

- [Sarawagi et al., 1998] Sarawagi, S., Thomas, S., and Agrawal, R. (1998). Integrating association rule mining with relational database systems: alternatives and implications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 343–354, New York, NY, USA. ACM.
- [Sarwar et al., 2001] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA. ACM.
- [Savova et al., 2010] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.
- [Sen et al., 2009a] Sen, S., Vig, J., and Riedl, J. (2009a). Learning to recognize valuable tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 87–96. ACM.
- [Sen et al., 2009b] Sen, S., Vig, J., and Riedl, J. (2009b). Tagommenders: connecting users to items through tags. In *Proceedings of the 18th international conference on World wide web*, pages 671–680. ACM.
- [Shepitsen et al., 2008] Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266. ACM.
- [Sollenborn and Funk, 2002] Sollenborn, M. and Funk, P. (2002). Category-Based Filtering and User Stereotype Cases to Reduce the Latency Problem in Recommender Systems. In Craw, S. and Preece, A., editors, *Advances in Case-Based Reasoning*, number 2416 in Lecture Notes in Computer Science, pages 395–405. Springer Berlin Heidelberg.
- [Spitzer et al., 1993] Spitzer, M., Braun, U., Hermle, L., and Maier, S. (1993). Associative semantic network dysfunction in thought-disordered

- schizophrenic patients: Direct evidence from indirect semantic priming. *Biological Psychiatry*, 34(12):864–877.
- [Steck, 2011] Steck, H. (2011). Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 125–132. ACM.
- [Straus and Haynes, 2009] Straus, S. and Haynes, R. (2009). Managing evidence-based knowledge: the need for reliable, relevant and readable resources. *Canadian Medical Association Journal*, 180(9):942–945.
- [Suchal and Návrat, 2010] Suchal, J. and Návrat, P. (2010). *Full Text Search Engine as Scalable k-Nearest Neighbor Recommendation System*, pages 165–173. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Sujatha and Dhavachelvan, 2011] Sujatha, P. and Dhavachelvan, P. (2011). Precision at K in Multilingual Information Retrieval. *International Journal of Computer Applications*, 24(9):40–43.
- [Suliman et al., 2016] Suliman, D., Malek, M., Kadima, H., and Laurent, D. (2016). Toward social-semantic recommender systems. *International Journal of Information Systems and Social Change (IJISSC)*, 7(1):1–30.
- [Symeonidis, 2016] Symeonidis, P. (2016). *Matrix and Tensor Factorization with Recommender System Applications*, volume 39. CRC Press.
- [Symeonidis et al., 2008] Symeonidis, P., Nanopoulos, A., Papadopoulos, A. N., and Manolopoulos, Y. (2008). Collaborative recommender systems: Combining effectiveness and efficiency. *Expert Systems with Applications*, 34(4):2995–3013.
- [Sánchez-Bocanegra et al., 2013] Sánchez-Bocanegra, C., Rivero-Rodríguez, A., Fernández-Luque, L., and Sevillano, J. L. (2013). Diavideos: a diabetes health video portal. In Lehmann, C. U., Ammenwerth, E., and Nøhr, C., editors, *MedInfo*, volume 192 of *Studies in Health Technology and Informatics*, page 941. IOS Press.
- [Sánchez Bocanegra, 2011] Sánchez Bocanegra, C. L. (2011). Rare Disease Video Portal. [http://openaccess.uoc.edu/webapps/o2/handle/10609/8046?mode=full&submit\\_simple>Show+full+item+record](http://openaccess.uoc.edu/webapps/o2/handle/10609/8046?mode=full&submit_simple>Show+full+item+record).



- [Sánchez-Bocanegra et al., 2012] Sánchez-Bocanegra, C. L., Rivero, A., Femandez-Luque, L., and Sevillano, J. L. (2012). Salutedca, a spanish health video portal. *Proceedings IV Workshop on Technology for Healthcare and Healthy Lifestyle*.
- [Sánchez-Bocanegra et al., 2015] Sánchez-Bocanegra, C. L., Sanchez-Laguna, F., and Sevillano, J. L. (2015). *Introduction on health recommender systems*, volume 1246. Springer.
- [Tan, 2005] Tan, S. (2005). Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28(4):667 – 671.
- [Teodoro et al., 2012] Teodoro, D., Pasche, E., Gobeill, J., Emonet, S., and Ruch, Patrick and Lovis, C. (2012). Building a transnational biosurveillance network using semantic web technologies: Requirements, design, and preliminary evaluation. *J Med Internet Res*, 14(3):e73.
- [Tomanek et al., 2007] Tomanek, K., Wermter, J., and Hahn, U. (2007). Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the Linguistic Annotation Workshop, LAW '07*, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Turner et al., 2015] Turner, M. A., Dew, N. K., Desai, L., Martin, N., and Kirchoff, K. (2015). Machine translation of public health materials from english to chinese: A feasibility study. *JMIR Public Health Surveill*, 1(2):e17.
- [Valizadegan et al., 2009] Valizadegan, H., Jin, R., Zhang, R., and Mao, J. (2009). Learning to rank by optimizing ndcg measure. In *Advances in neural information processing systems*, pages 1883–1891.
- [Van de Belt et al., 2012] Van de Belt, T. H., Berben, S. A. A., Samsom, M., Engelen, L. J. L. P. G., and Schoonhoven, L. (2012). Use of social media by western european hospitals: longitudinal study. 14(3):e61.
- [Vance et al., 2009] Vance, K., Howe, W., and Dellavalle, R. P. (2009). Social internet sites as a source of public health information. *Dermatologic Clinics*, 27(2):133 – 136. *Dermatologic Epidemiology and Public Health*.

- [Ventola, 2014] Ventola, C. L. (2014). Social media and health care professionals: Benefits, risks, and best practices. *Pharmacy & Therapeutics (P T)*, 39(7):491–520.
- [Vig et al., 2009] Vig, J., Sen, S., and Riedl, J. (2009). Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56. ACM.
- [Vig et al., 2010] Vig, J., Soukup, M., Sen, S., and Riedl, J. (2010). Tag expression: tagging with feeling. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 323–332. ACM.
- [von Niman et al., 2006] von Niman, B., Rodriguez-Ascaso, A., Sund, T., and Brown, S. J. (2006). User experience design guidelines for telecare services. In *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services, MobileHCI '06*, pages 279–280, New York, NY, USA. ACM.
- [Wang and Ye, 2009] Wang, P. and Ye, H. (2009). A Personalized Recommendation Algorithm Combining Slope One Scheme and User Based Collaborative Filtering. In *International Conference on Industrial and Information Systems, 2009. IIS '09*, pages 152–154.
- [Wang et al., 2013] Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*.
- [Wiesner and Pfeifer, 2010] Wiesner, M. and Pfeifer, D. (2010). Adapting recommender systems to the requirements of personal health record systems. In *Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10*, pages 410–414, New York, NY, USA. ACM.
- [Wiljer et al., 2008] Wiljer, D., Urowitz, S., Apatu, E., DeLenardo, C., Eysenbach, G., Harth, T., Pai, H., and Leonard, J. K. (2008). Patient accessible electronic health records: Exploring recommendations for successful implementation strategies. *J Med Internet Res*, 10(4):e34.

- [Wilson et al., 2016] Wilson, A., Childs, S., Gonçalves-Bradley, D., and Irving, G. (2016). Interventions to increase or decrease the length of primary care physicians’ consultation. *Cochrane Database Syst Rev*.
- [Wong and Hong, 2007] Wong, J. and Hong, J. I. (2007). Making mashups with marmite: Towards end-user programming for the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’07, pages 1435–1444, New York, NY, USA. ACM.
- [Wray and Eklund, 2010] Wray, T. and Eklund, P. W. (2010). Social tagging for digital libraries using formal concept analysis.
- [Yokoo et al., 1998] Yokoo, M., Durfee, E. H., Ishida, T., and Kuwabara, K. (1998). The distributed constraint satisfaction problem: Formalization and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 10:673–685.
- [Zeng et al., 2006] Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., and Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90.
- [Zhang et al., 2006] Zhang, S., Ouyang, Y., Ford, J., and Makedon, F. (2006). Analysis of a low-dimensional linear model under recommendation attacks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 517–524. ACM.
- [Zhang et al., 2012] Zhang, Y. C., Séaghdha, D. O., Quercia, D., and Jambor, T. (2012). Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM ’12, pages 13–22, New York, NY, USA. ACM.
- [Zhao, 2015] Zhao, Y. (2015). Association rule mining with r.
- [Žitnik and Zupan, 2015] Žitnik, M. and Zupan, B. (2015). Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53.