xx Análisis de Redes Sociales
Teoría de Grafos
xx Aprendizaje Automático
Procesos de Markov
xx Sistemas de Recomendación
Procesamiento de Lenguaje Natural

Optimización de Recomendación de Información Basada en Técnicas de Procesamiento de Lenguaje Natural y Análisis de Redes Sociales

Autor: Pablo Gabriel Celayes.
Director: Dr. Martín A. Domínguez.
Facultad de Matemática, Astronomía y Física — U.N.C.

29 de septiembre de 2015

Resumen

Dada la gran cantidad de información disponible actualmente en Internet y su velocidad de generación, se vuelve cada vez más difícil y tedioso encontrar contenido actualizado y de interés. Se genera entonces la necesidad de contar con aplicaciones que faciliten la clasificación de la información y el filtrado de artículos de valor informativo para cada usuario. El proyecto en desarrollo ?Cogfor? (donde el alumno se desempeña desde hace unos meses como programador/investigador) viene a proveer una solución a este problema tanto para entornos corporativo como para uso personal. La presente tesis se centra en hacer un aporte a la inteligencia de filtrado de dicho proyecto.

Actualmente la plataforma mencionada evalúa el contenido consumido por una organización diariamente, con el fin de clasificarlo y seleccionarlo para los integrantes de acuerdo a sus preferencias personales. De momento, las recomendaciones generadas para cada usuario se basan solamente en su uso de la plataforma, sin tener en cuenta al resto de los usuarios.

El aporte de la presente tesis será agregar una dimensión social al proceso, modelando no sólo las preferencias de cada usuario, sino además sus afinidades con otros usuarios y comunidades a las que pertenece. Esto nos da la oportunidad de combinar dos áreas de investigación muy dinámicas y actualmente en boga, como son el Procesamiento de Lenguaje Natural (PLN) y el Análisis de Redes Sociales (ARS).

Un paso inicial en la integración de estas dos áreas es resolver el problema de ?cold start?, que consiste en dar un modelo inicial para las preferencias de un usuario nuevo, sobre el que todavía no tenemos comportamiento registrado. Las técnicas de ARS nos permitirán descubrir usuarios existentes con gustos similares al nuevo, para así basar las sugerencias iniciales en lo ya conocido sobre dichos usuarios.

A continuación, se investigarán métodos para mejora continua de recomendaciones basada en información sobre la afinidad y conectividad entre usuarios. Se estudiarán diversas maneras de extraer dicha información de fuentes externas (redes sociales como Facebook, LinkedIn o Twitter) o internas (interacciones por mail en el caso corporativo) empleando técnicas de detección de comunidades para encontrar tanto temas de interés como usuarios afines en quienes basar nueva recomendaciones.

TODO: adaptar

II

Agradecimientos

En ningún orden particular:

A todos.

Índice general

Capítulo 1

Introducción

TODO: adaptar

La teoría de LMP (Labelled Markov Processes) desarrollada en [99Des] se ocupa del análisis de sistemas probabilistas sobre un espacio de estados continuo, e introduce una noción de equivalencia entre tales sistemas —la relación de bisimulación, inspirada en la versión discreta de [91LS]— que formaliza el concepto de que dos sistemas tengan el mismo "comportamiento observable". La tarea de analizar y razonar sobre un sistema, se hace mucho más precisa si se establecen métodos formales a tal fin. Un método formal es un conjunto de lenguajes, técnicas y herramientas de gran rigor matemático, empleado para especificar (describir) y verificar sistemas. Un objetivo general en esta área es abordar el estudio de las propiedes observables de los LMP mediante métodos formales. Como primer paso en esta dirección, se probó que la bisimulación está caracterizada por una lógica modal simple $\mathcal{L}_{>}$ (sin negación), siendo equivalentes sólo aquellos sistemas con las mismas propiedades definibles en dicha lógica. Para demostrar este resultado fueron esenciales algunos teoremas y técnicas de teoría de la medida, junto con propiedades de los espacios analíticos.

Tomando esto como motivación, estudiamos los conceptos básicos de teoría descriptiva de conjuntos necesarios para llegar a demostrar toda la maquinaria matemática empleada en [99Des], y comprender con más profundidad las razones que llevaron a emplear espacios analíticos en la definición. A partir de ahí, buscamos extender los resultados de caracterización lógica a procesos no deterministas (los NLMP, Non-deterministic Labelled Markov Processes), partiendo de las definiciones básicas de [06W-D'A]. Veremos que la teoría de NLMP y LMP puede construirse tomando sólo espacios Polacos como conjuntos de estados, que si bien son menos generales que los analíticos, tienen una definición menos técnica y un poco más cercana a la intuición. Esto es posible si definimos las relaciones sobre un mismo sistema (prescindiendo del enfoque categórico), lo que evita trabajar con espacios cociente (una de las principales razones para emplear espacios analíticos de estados en [99Des]). Reconstruimos el teorema de caracterización para nuestras definiciones, e introducimos un conjunto de lógicas modales más expresivas para varias familias cada vez

más amplias de NLMP (llegando hasta los NLMP de imagen finita, aquellos con grados finitos arbitrarios de no determinismo).

El capítulo 2 abarca los contenidos de teoría descriptiva de conjuntos y teoría de la medida. Allí definimos las herramientas básicas (espacios métricos, árboles, espacios de Cantor y de Baire) y realizamos un breve estudio de los espacios Polacos, y sus subconjuntos Borel y analíticos. Es de fundamental importancia en la teoría de NLMP el Teorema de Separación de Lusin, y sus aplicaciones a las relaciones de equivalencia Borel (Teorema de Blackwell) y a la medibilidad de ciertas proyecciones (Lusin-Novikov). Al final del capítulo probamos dos resultados útiles sobre coincidencia de medidas. El primero de ellos aparecía sin prueba en [99Des], el segundo es un lema que probamos para poder caracterizar la bisimulación en 2-NLMP por una lógica simple.

En el capítulo **??**, partiendo de [06W-D'A], se introducen las definiciones básicas de NLMP y la relación de bisimulación sobre ellos, junto con las pruebas de algunas propiedades de esta relación. Se incluye una sección de "interfaz con Desharnais", que muestra como adaptar nuestras definiciones a las de [99Des]. Definimos luego las lógicas modales a emplear en la caracterización. Por último se aborda la prueba de caracterización lógica, dando una prueba estructural que engloba todas aquellas propiedades que no dependen de la lógica usada. Una vez hecho esto, probar la caracterización para una lógica en particular se reducirá a probar que ésta satisface una cierta hipótesis de coincidencia de medidas. Concluimos aplicando lo anterior a la prueba de caraterización de la bisimulación para cuatro clases de NLMP por distintas lógicas modales.

Capítulo 2

Fundamentos Teóricos

2.1. Aprendizaje Automático

2.1.1. Clasificación

2.2. Análisis de Redes Sociales

2.2.1. Medidas de afinidad entre nodos

2.2.2. Detección de comunidades

2.3. Procesamiento de Lenguaje Natural

2.3.1. Topic Modeling

Capítulo 3

Datos de muestra

3.1. Idea original: usuarios propios

3.2. Plan B: Escrapear Facebook

3.3. Plan C: Twitter API

3.3.1. Construcción de grafo de muestra

3.3.2. Recolección de tweets

3.4. Problemática de los Sistemas de Recomendación

3.5. Evaluación y comparación de técnicas

Capítulo 4

Conclusiones y Trabajo Futuro

TODO: adaptar

El principal aporte de este trabajo es la construcción de la teoría de NLMP sobre espacios Polacos en vez de analíticos, abriendo una línea que no aparecía como posible en [99Des]. Si bien los espacios analíticos son más generales que los Polacos, su definición es más técnica, con las complicaciones que ello conlleva. Notemos que los espacios analíticos no fueron introducidos para darle más generalidad a la teoría sino por una limitación técnica que impedía usar sólo espacios Polacos (debido al uso de cocientes en la prueba). En la práctica, es difícil imaginar procesos que requieran el uso de un espacio analítico no Polaco como conjunto de estados.

El estudio de los espacios Polacos y analíticos y sus propiedades relevantes en la teoría de NLMP, permitió lograr, además del aporte ya mencionado, una mejor estructuración de las pruebas de caracterización lógica de la bisimulación, aislando una propiedad central que garantiza que una lógica caracteriza la bisimulación. Notemos que los conjuntos analíticos, a pesar de ya no ser usados como espacios de estados, siguen teniendo un rol fundamental en la demostración de los resultados técnicos utilizados (los teoremas de Blackwell y de Lusin-Novikov, son ambos consecuencias del Teorema de Separación de Lusin).

Sin embargo, puede ser interesante ver que nuestros resultados se verifican también sobre espacios analíticos de estados, ya que esto permitiría una mejor integración con los resultados de [99Des] y todo el desarrollo subsiguiente. También se podría intentar dar una organización categórica de los NLMP. Otras posibles mejoras se refieren a la eliminación de algunas de las restricciones impuestas aquí, como por ejemplo, retirar la condición de imagen finita o trabajar sobre procesos con conjuntos no numerables de acciones (por ejemplo, en los que la interacción con el entorno pueda estar dada por variables temporales o físicas continuas).

Bibliografía

[06W-D'A] Nicolás Wolovick, Pedro D'Argenio: Trabajo en progreso. *FaMAF, Córdoba*

[05Des] Vincent Danos, Josée Desharnais, François Laviolette, Prakash Panangaden: "Bisimulation and Cocongruence for Probabilistic Systems". *McGill University, Montréal, Québec - Université Paris 7 and CNRS - Université Laval, Québec, 2005*

[01vanG] Rob J. van Glabbeek: "The linear time - branching time spectrum I; the semantics of concrete, sequential processes", en: "Handbook of Process Algebra". *Chapter 1, Elsevier, 2001, pp. 3-99*

[99Des] Josée Desharnais: "Labelled Markov Processes" *School of Computer Science. McGill University, Montréal, 1999*

[96CrHu] Max J. Cresswell, G. E. Hughes: "A New Introduction to Modal Logic" *Routledge, London, 1996*

[95Kec] Alexander S Kechris: "Classical Descriptive Set Theory". *Springer-Verlag, 1995*

[91LS] K.G. Larsen, A. Skou: "Bisimulation through Probabilistic Testing" en "Information and Computation", *94(1):1-28, 1991*

[89Mil] R.Milner: "Communication and Concurrency"*Prentice Hall, 1989*

[86Bil] Patrick Billingsley: "Probability and Measure". *John Wiley and Sons, New York, 1986*

[70Rud] Walter Rudin: "Real and Complex Analysis". *McGraw Hill, 1970*

[Wiki] Wikipedia `http://wikipedia.org`