

clínica de pacientes (Afantenos, Karkaletsis y Stamatopoulos, 2005).

Si a ello unimos el hecho de que gran parte de los resultados de la investigación biomédica se encuentran en forma de literatura escrita en formato libre (no estructurados, formato inadecuado para la búsqueda compleja) que se acumulan en grandes bases de datos en línea, podemos concluir que el proceso de reducción de los resúmenes automáticos es especialmente útil en el ámbito biomédico.

Por otro lado, el rápido crecimiento de los resultados de la investigación del dominio biomédico está produciendo un importante cuello de botella. MEDLINE (*Medical Literature Analysis and Retrieval System Online*), la principal base de datos bibliográfica de EE.UU. (de la *National Library of Medicine*), contiene más de 16 millones de referencias a artículos de revistas, centrados principalmente en biomedicina. Entre 2000 y 4000 referencias completas se añaden cada día, más de 670000 fueron añadidas en 2007².

La práctica de la medicina basada en la evidencia ha sido tradicionalmente definida como la combinación de los mejores resultados de la investigación médica con el juicio clínico, experto y experimentado (Sackett et Al., 1996). La capacidad de buscar en la literatura médica en un tiempo eficiente representa una parte importante de una práctica basada en la evidencia. Un reciente trabajo cualitativo concluyó que dos de los seis obstáculos para responder a cuestiones clínicas aplicando la evidencia eran el tiempo requerido para encontrar información y la dificultad para seleccionar una estrategia óptima de búsqueda (Ely y Osheroff, 2002). Es por todo esto que herramientas de búsqueda como PubMed³, BioMed Central⁴ o UpToDate⁵ se han convertido en más y más importantes, para encontrar formas adecuadas de localizar la mejor evidencia de manera eficaz.

En este dominio, los profesionales en general necesitan herramientas orientadas a proporcionar medios para acceder y visualizar la información adecuada para sus necesidades.

En este trabajo presentamos un modelo de generación de resúmenes de carácter extractivo apoyado en conceptos del dominio biomédico. El artículo se estructura de la siguiente manera: en primer lugar se describe el proceso de tratamiento extractivo del lenguaje natural mediante el uso de grafos, para posteriormente comentar algunos trabajos específicos del dominio. Presentamos UMLS y el conjunto de herramientas de procesamiento de lenguaje natural orientadas al ámbito biomédico que incorpora. En la sección cinco presentamos el modelo de generación de resúmenes en que estamos trabajando, dividido en cuatro fases: la generación del grafo léxico, la aplicación de un algoritmo de similitud conceptual, la aplicación de un algoritmo de ranking y finalmente, la creación del resumen. Finalmente enumeramos los muchos temas abiertos que quedan en este trabajo inicial y los posibles futuros trabajos.

2 Trabajos relacionados en el ámbito extractivo

Para generar resúmenes automáticos de texto existen dos enfoques: extractivo y abstractivo. El enfoque extractivo selecciona y extrae frases o partes de ella del texto original. La mayor ventaja que tiene este enfoque es que resulta muy robusto y fácilmente aplicable a contextos de propósito general, ya que, su independencia del dominio, e incluso del género de los documentos, es muy alta. El enfoque abstractivo suele englobar técnicas de procesamiento del lenguaje natural, más complejo pues necesita un conocimiento léxico, gramatical y sintáctico del dominio, para modelar semánticamente el conocimiento y a partir de éste ser capaz de generar un resumen.

Típicamente, el proceso de resumen extractivo consiste en identificar las sentencias de un texto de origen que sean relevantes para el usuario a la vez que se reduce la redundancia de la información. Las sentencias son puntuadas basándose en una serie de características y las n sentencias de mayor puntuación son extraídas y presentadas al usuario en su orden de aparición en el texto original.

Para trabajar con las frases y su puntuación, un mecanismo de representación comúnmente usado han sido los modelos de puntuación o *ranking* basados en grafos. Los algoritmos de

² <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

³ <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>

⁴ <http://www.biomedcentral.com/info/>

⁵ <http://www.uptodate.com/home/about/index.html>

ranking basados en grafos son un modo de decidir sobre la importancia de un vértice dentro del grafo, teniendo en cuenta información referencial global del grafo, obtenida recursivamente mejor que localmente desde el vértice.

La aplicación de éste modo de trabajo a grafos léxicos o semánticos extraídos de documentos de lenguaje natural ha sido llevada a cabo (Skorochoďko, 1972) (Salton et al., 1997) y se ha mostrado eficaz en tareas de procesamiento del lenguaje como la extracción automática de palabras clave, generación de resúmenes extractiva o desambiguación del sentido de las palabras (Mihalcea y Tarau, 2006).

Otros trabajos relevantes en el ámbito que destacaremos son (Radev y McKeown, 1998) donde se presenta un sistema que genera un resumen a partir de un conjunto de artículos periodísticos sobre el mismo acontecimiento. Para cada frase se determina su estructura a alto nivel y las palabras que van a representar cada papel semántico y, finalmente, se construye su árbol sintáctico.

El sistema SUMMARIST (Hovy y Lin, 1999) se utiliza un recurso léxico, WordNet para identificar conceptos genéricos y definir una jerarquía. El proceso de generalización se realiza mediante la propagación de pesos de los conceptos, basados en frecuencias de aparición, a través de la jerarquía de WordNet.

3 Trabajos relacionados en el ámbito biomédico.

En el ámbito biomédico destacaremos los métodos de generación de resúmenes extractivos como BioChain, (basado en cadenas de conceptos o relaciones semánticas entre conceptos vecinos en texto), FreqDist (centrado en el uso de las distribuciones de frecuencia, construyendo un resumen con similar distribución que el original) y Chainfreq (híbrido de los dos anteriores), que usan conceptos específicos del dominio biomédico para identificar las sentencias destacables del texto completo (Reeve, Han y Brooks, 2007). Sin embargo, la posterior evaluación de los métodos no logra mejorar los resultados de los enfoques basados en términos.

Los trabajos específicos de un ámbito pueden usar conceptos en vez de términos, para lo que necesitan herramientas que den soporte a la identificación de los conceptos en una estructura de conocimiento del dominio y capaces de determinar relaciones semánticas entre estos conceptos.

3.1 Conocimiento del dominio: UMLS

Para el procesado semántico, consistente en el análisis e identificación de los conceptos y relaciones subyacentes en un texto, se requiere para que el texto pueda ser mapeado a una estructura de conocimiento, como la que en el ámbito biomédico proporciona el proyecto *Unified Medical Language System* (UMLS) (Humphreys et al., 1998). El objetivo de este proyecto es el desarrollo de herramientas que ayuden a investigadores en la representación del conocimiento, recuperación e integración de información biomédica.

UMLS consiste en tres componentes, el *SPECIALIST Lexicon*, el *Metathesaurus* y la *UMLS Semantic Network* (Rindflesh, Fiszman y Libbus, 2005).

- El *SPECIALIST Lexicon* describe las características sintácticas de terminos en inglés de carácter biomédico y general, proporcionando la base para el PLN en el dominio biomédico.

Así, p.ej., la entrada 'Anaesthetic' produciría las siguientes respuestas:

- *{base=anesthetic*
 - *spelling_variant=anaesthetic*
 - *entry=E0330018*
 - *cat=noun*
 - *variants=reg*
 - *variants=uncount }*
- *{base=anesthetic*
 - *spelling_variant=anaesthetic*
 - *entry=E0330019*
 - *cat=adj*
 - *variants=inv*
 - *position=attrib(3)*
 - *position=pred stative },*

que vendría a indicarnos que el término puede aparecer como sustantivo o adjetivo, en un caso con un plural regular, incontable, en el otro indica que es invariante, que puede aparecer en el predicado y que es un adj. atributivo.

- El *Metathesaurus* es una recopilación de más de 100 vocabularios y terminologías médicas, entre los que se incluyen desde MeSH o SNOMED hasta subdominios más especializados (odontología o enfermería,...) asociando cada término a más de un millón de conceptos semánticos que a su vez se engloban en 135 tipos semánticos relevantes en el ámbito biomédico (y siempre, al menos en uno).

Así, p.ej., la entrada 'Arthritis, Juvenile Rheumatoid' produciría la siguiente información jerárquica:

Immunologic Diseases

Autoimmune Diseases

Arthritis, Rheumatoid

Arthritis, Juvenile Rheumatoid

- La *UMLS Semantic Network* constituye una ontología del más alto nivel de la Medicina, compuesta por 135 tipos semánticos asignados a conceptos del *Metathesaurus* y por 54 tipos de relaciones entre los tipos. Estas relaciones son a menudo llamadas predicados o proposiciones y están constituidas por argumentos (conceptos) y predicados (relaciones). Algunos ejemplos podrían ser:

- 'Therapeutic or Preventive Procedure'

TREATS 'Injury or Poisoning'

- 'Organism Attribute' PROPERTY_OF 'Mammal'

- 'Bacterium' CAUSES 'Pathologic Function'.

SemRep es una herramienta de procesamiento semántico que integra los tres anteriores componentes de UMLS para analizar de manera automática textos con lenguaje médico identificando los conceptos y relaciones que representan el contenido del documento. SemRep devuelve una lista de relaciones a partir de un conjunto de documentos obtenidos por una búsqueda de un término especificado.

Usaremos el *Metathesaurus* y la herramienta *Metamap Transfer* (MMTx) para la identificación de los conceptos biomédicos de cada frase, base para el cálculo del solape entre frases. En cuanto a SemRep, añadiremos esta lista de relaciones al grafo dirigido para posteriores trabajos.

En castellano han existido esfuerzos para la elaboración de un *metathesaurus*, como WordMed (Arranz et al., 2000). Destacaremos el trabajo de (Carrero, Cortizo y Gómez, 2008)

que combina técnicas de traducción automática con ontologías biomédicas y MMTx para producir una versión española de MMTx.

4 Propuesta de generación del resumen

Los métodos de generación de resúmenes basados en técnicas extractivas han demostrado ser muy útiles por su adaptabilidad y eficiencia en tiempo de respuesta en cualquier tipo de dominios. Por contra, los métodos abstractivos, por la necesidad de recursos léxicos, sintácticos y semánticos han proporcionado unos mejores resultados en cuanto a comprensibilidad a costa de un mayor esfuerzo computacional y por tanto, de tiempos de respuesta, aparte de la especificidad del ámbito de uso de la herramienta.

Como vimos en el punto dos, existen trabajos previos para el dominio específico biomédico de carácter extractivo que hacen uso de recursos léxicos y semánticos, pero que no obtienen unos mejores resultados trabajando con conceptos que con términos. Nuestro objetivo es intentar mejorar la capacidad y rapidez de los métodos extractivos con la efectividad y concreción de los métodos abstractivos. Para ello vamos a presentar una primera propuesta de una metodología de generación automática de resúmenes basada en conocimiento estructurado y grafos de *ranking*.

Nuestra propuesta, basada en (Mihalcea y Tarau, 2006) es eminentemente extractiva, de modo que el proceso podría resumirse en identificar las sentencias en el texto de origen, seleccionar aquellas que sean relevantes para el usuario a la vez que disminuimos la redundancia de la información. Para ello asignamos una puntuación a cada frase de acuerdo a un conjunto de características. Las *n*-primeras frases en cuanto a puntuación se extraen y se presentan al usuario en su orden de aparición en el texto original.

4.1 Fase 1. Generación del grafo.

Independientemente del tamaño del texto, sea un texto completo o un *abstract*, la primera tarea debe consistir en la identificación de cada una de las sentencias del texto de origen, así como en la creación de un grafo que incluya un vértice en el grafo por cada sentencia. De manera simultánea, se identifican con la ayuda

de *Metamap Transfer* (integrada en SemRep, ver *Figura 1*), los conceptos biomédicos incluidos en la frase y se incluyen en el nodo, así como las relaciones semánticas. Para el trabajo con grafos en el prototipo que se ha elaborado se ha usado la librería JUNG (O'Madadhain et al., 2004).

```
SE|00000000||tx|1|text|In order to
substantiate further the relationship between
these oral disorders and psoriasis, we
compared 200 patients with psoriasis to a
matched control group.
SE|00000000||tx|1|entity|C1517331|Further|spco
||further|||888|26|32
SE|00000000||tx|1|entity|C0439849|Relationship
s|qlco|||relationship|||888|38|49
SE|00000000||tx|1|entity|C0026636|Mouth
Diseases|dsyn||oral disorders|||983|65|78
SE|00000000||tx|1|entity|C0033860|Psoriasis|ds
yn||psoriasis|||1000|84|92
SE|00000000||tx|1|entity|C0030705|Patients|pod
g||patients|||861|111|118
SE|00000000||tx|1|entity|C0033860|Psoriasis|ds
yn||psoriasis|||1000|125|133
SE|00000000||tx|1|entity|C0243148|control|ftcn
||control|||901|148|154
SE|00000000||tx|1|entity|C0024908|Matched
Groups|grup||matched control
group|||901|140|160
SE|00000000||tx|1|relation|2|1|C0033860|Psoria
sis|dsyn|dsyn||psoriasis|||1000|125|133|PREP
|PROCESS_OF|120|123|5|1|C0030705|Patients|hmn
n|humn||patients|||861|111|118
```

Figura 1 Ejemplo de tratamiento realizado por SemRep sobre una sentencia de un texto biomédico

4.2 Fase 2. Aplicación de algoritmo de similitud.

Para la extracción de sentencias en resúmenes, un concepto importante es la 'similitud' o grado de solapamiento entre sentencias, cuánto del contenido de una sentencia se encuentra incluido en otra. Es como si consideráramos el solape como una "recomendación" de una frase de dirigirse a otras que tratan y abundan los mismos conceptos. Una función de similitud, que tome en cuenta el grado de repetición de *tokens* entre sentencias de manera normalizada proporcionará una medida de este concepto. En particular, este concepto también nos proporcionará información de lo cohesionado o no del grupo de documentos devueltos en la consulta y de la posible necesidad de un tratamiento previo de *clustering*.

Aplicamos una versión modificada (con conceptos en vez de términos) de la formula de similitud de (Milhacea y Tarau, 2006):

$$Similitud_c(V_i, V_j) = \frac{(|C_k / C_k \in V_i \wedge C_k \in V_j|)}{\log((V_i)) + \log((V_j))}$$

La Figura 2 muestra la matriz de adyacencia que almacena los pesos de las aristas entre nodos, así de cada nodo en una fila a un nodo de una columna (grafo dirigido) se muestra en la tabla el valor de similitud.

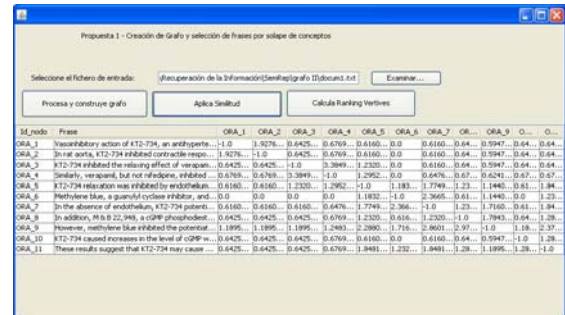


Figura 2: Prototipo de la aplicación tras aplicar algoritmo de solape

4.3 Fase 3. Aplicación de algoritmo de ranking

Los algoritmos de *ranking* basados en grafos, a partir de la asignación arbitraria de valores a cada nodo, realizan cálculos para obtener la puntuación $S(V_i)$ de cada nodo de manera iterativa, hasta que se produce convergencia bajo un determinado umbral. Las referencias entre nodos y/o conceptos son tratadas como 'votos' para decidir el elemento más importante. La puntuación de cada vértice se obtiene aplicando PageRank (Brin y Page, 1998):

$$WS(V_i) = (1 - d) + d * \sum_{v_j \in I_n(V_i)} \frac{W_{ji}}{\sum_{v_k \in O_{ut}(V_j)} W_{jk}} WS(V_j)$$

En la Figura 3 se observa el prototipo de la aplicación con el grafo resultante, donde se pueden observar los nodos etiquetados con los pesos obtenidos y los valores asociados a las aristas recalculados.

Tras la ejecución del algoritmo, los nodos se ordenan atendiendo al peso o puntuación asociada, que define la notoriedad (*saliency*) de cada vértice en un grafo dirigido y ponderado.

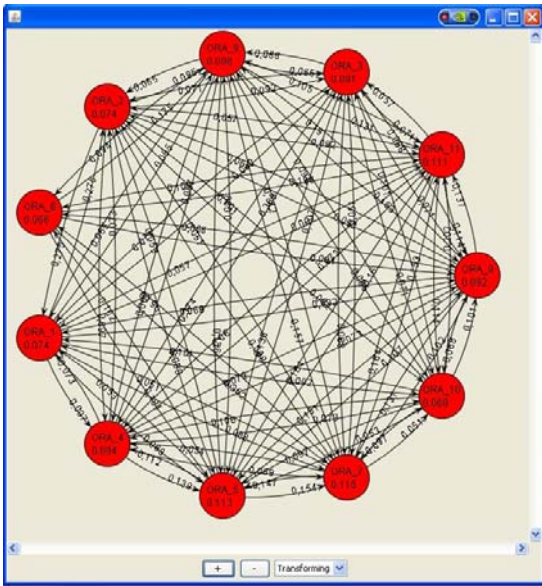


Figura 3: Gráfico del Grafo con pesos generado por algoritmo de ranking

4.4 Fase 4. Creación del resumen

Los nodos de mayor puntuación definirán las frases a incluir en el resumen. El número de frases puede ser fijo o basado en umbral o porcentaje. En nuestro prototipo es el usuario el que decide el porcentaje de frases.

Para facilitar la legibilidad del resumen, la secuencialidad de presentación de las frases seleccionadas se hace atendiendo a su ordenamiento original.

5 Conclusión y temas abiertos

Se ha presentado una propuesta de generación automática de resúmenes de carácter extractivo, que usa una representación en grafo donde los nodos son frases y las aristas un valor numérico que mide el 'grado de recomendación' o similitud entre frases. El algoritmo de ranking producirá como resultado un peso en los nodos, que representa la importancia global de la frase dentro del documento, que ordenaremos de mayor a menor. Seleccionaremos las primeras en un número determinado por el porcentaje de compresión indicado a la herramienta.

La novedad de la metodología se encuentra en el uso del metatesauro UMLS para identificar conceptos UMLS y que la similitud entre frases se calcule a partir del número de conceptos UMLS que compartan las frases. Entendemos que la herramienta aún las bondades de técnicas extractivas con el conocimiento del dominio que aportan los recursos UMLS y que

debe reflejarse en un buen resultado en una futura evaluación de método.

Es evidente que la propuesta es un punto de partida que acabará como un hito en un proyecto más ambicioso y a más largo plazo. Hablemos de cuáles serán los siguientes pasos a realizar:

- Elaboración u obtención de un corpus evaluable. En este momento nos encontramos en la búsqueda de un corpus que podamos reutilizar para nuestros fines. De no tener un resultado positivo, optaríamos por elaborar nuestro propio corpus de documentos, a partir de BioMed Central, una editorial independiente dedicada a la publicación de artículos de investigación en Biología y Medicina que se caracteriza por mantener una política de acceso abierto a través de Internet, agrupando a más de 180 revistas y más de 23000 artículos de investigación del ámbito biomédico. Esto nos permitiría trabajar con un amplio conjunto de documentos completos en vez de *abstracts*.
- Evaluación. Cualquier trabajo mínimamente metódico requiere de una comparación de su eficiencia frente a otras propuestas de prestigio y frente a un *baseline* que proporcione métricas sobre los porcentajes de mejora por aplicación de tal o cual modificación. Nos proponemos evaluar nuestro modelo usando uno de estas herramientas:
 - ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin y Hovy, 2003) es una herramienta automatizada que compara un sumario generado por un sistema automático con uno o más resúmenes ideales, llamados modelos. Usa N-gramas para determinar el solape entre el resumen generado y los modelos.
 - Basic Elements (Hovy et al., 2006) es un marco de trabajo en el que las medidas de evaluación de los resúmenes pueden instanciarse y compararse dentro de un método de evaluación que se basa en el trabajo con unidades de contenido muy pequeñas, llamados 'basic elements' que corrigen algunos de los defectos de los n-gramas.

Parece lógico que la segunda herramienta, basada en la comparación de pequeñas unidades de contenido en vez de n-gramas, favorecerá a una herramienta basada en conceptos en vez de en cadenas. Sin embargo, actualmente BE no se encuentra soportado.

- Nos planteamos la evolución y mejora de esta propuesta analizando y haciendo uso de las relaciones semánticas obtenidas mediante SemRep. Nuestra idea es incluirlas dentro del grafo, de modo que dos conceptos unidos mediante una relación generarán una arista dirigida entre los nodos que incluyan a cada uno de esos conceptos. El peso de cada arista vendrá definido por el tipo de relación semántica (una relación 'cause' o 'threats' será más relevante que otra 'is-a').

Bibliografía

- Afantenos, S. D., V. Karkaletsis y P. Stamatoopoulos. 2005. Summarization from Medical Documents: A Survey en *Artificial Intelligence in Medicine*, 33(2):157-177.
- Arranz V., X. Carreras, M. A. Martí, J. Turmo, J. Vilalta. 2000. WORDMED: Un recurso conceptual terminológico para el desarrollo de aplicaciones de PLN en el dominio médico. *VII Simpósio Ibero-Americano de Terminologia: Terminologia e Indústrias da Língua*, Lisboa, (Portugal), noviembre de 2000.
- Brin, S. y L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30 (1-7). 1998.
- Ely, J.W., J.A. Osheroff, M.H. Ebell, M.L. Chambliss, D.C. Vinson, J.J. Stevermer y E.A. Pifer. 2002. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *British Medical Journal*, 324: 710.
- Carrero F.M., J.C. Cortizo y J.M. Gómez. 2008. Building a Spanish MMTx by Using Automatic Translation and Biomedical Ontologies. *IDEAL 2008*: 346-353
- Hovy, E. y C.Y. Lin. 1999. Automated Text Summarization in SUMMARIST. En I. Mani y M. T. Maybury, eds., *Advances in Automatic Text Summarization*, pags. 81-94. The MIT Press. 1999.
- Hovy, E., C. Y. Lin, L. Zhou, J. Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. En *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italia.
- Humphreys, B.L., D.A. Lindberg, H.M. Schoolman y G.O. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5(1), 1-11. 1998.
- Lin, C. Y. y E. Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. En *Proceedings of 2003 language technology conference (HLT-NAACL 2003)* (Vol. 1(1), pag. 71-78). Edmonton, Canada.
- Mihalcea R. y P. Tarau. 2006. TextRank: Bringing Order into Texts. En *Proceedings of Empirical Methods in Natural Language Processing*. ACL, 404-411, 2006.
- O'Madadhain, J., S. White, D. Fisher y Y. B. Boey. 2004. JUNG-Java Universal Network/graph Framework. Available for download at <http://jung.sourceforge.net/>.
- Radev, D. R. y K. R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 4:469-500.
- Reeve, L.H., H. Han, A.D. Brooks. 2007. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management* 43, 1765-1776. 2007.
- Rindfleisch, T.C., M. Fiszman, B. Libbus. 2005. Semantic interpretation for the biomedical research literature. Capítulo 14 del libro *Medical Informatics. Knowledge Management and Data Mining in Biomedicine* (Springer's Integrated Series in Information Systems), editores Chen, H., Fuller, S.S., Friedman C., Hersh, W.
- Sackett D.L., W.M.C. Rosenberg, J.A.M. Gray, R.B. Haynes y W.S. Richardson. 1996. Evidence-based medicine: what it is and what it isn't. *British Medical Journal*, 312: 71-72.
- Salton, G., A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management* 33 (3), 193-207.

Skorochod'ko, E. F. 1972. Adaptive method of automatic abstracting and indexing. En C. Freiman, ed., *Information Processing 71: Proceedings of the IFIP Congress 71*, págs.1179-1182. North-Holland Publishing Company, Amsterdam.