

# University of Wollongong - Research Online

## Thesis Collection

Title: Intention-driven textual semantic analysis

Author: Jie Li

Year: 2008

Repository DOI:

### Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

**Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.**

Research Online is the open access repository for the University of Wollongong. For further information contact the UOW Library: [research-pubs@uow.edu.au](mailto:research-pubs@uow.edu.au)

2008

## Intention-driven textual semantic analysis

Jie Li  
*University of Wollongong*

Follow this and additional works at: <https://ro.uow.edu.au/theses>

**University of Wollongong**

**Copyright Warning**

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

---

### Recommended Citation

Li, Jie, Intention-driven textual semantic analysis, MCompSc-Res thesis, School of Computer Science and Software Engineering, University of Wollongong, 2008. <http://ro.uow.edu.au/theses/104>

## **NOTE**

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

## **UNIVERSITY OF WOLLONGONG**

### **COPYRIGHT WARNING**

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.



# Intention-driven Textual Semantic Analysis

A thesis submitted in partial fulfillment of the  
requirements for the award of the degree

**Master of Computer Science**

from

UNIVERSITY OF WOLLONGONG

by

**Jie Li**

School of Computer Science and Software Engineering

July 2008

© Copyright 2008

All Rights Reserved

*Dedicated to*

*This thesis is dedicated to my mom and dad.*

# Declaration

This is to certify that the work reported in this thesis was done by the author, unless specified otherwise, and that no part of it has been submitted in a thesis to any other university or similar institution.

---

Jie Li  
July 18, 2008

# Abstract

---

The explosion of World Wide Web has brought endless amount of information within our reach. In order to take advantage of this phenomenon, text search becomes a major contemporary research challenge. Due to the nature of the Web, assisting users to find desired information is still a challenging task. In this thesis, we investigate semantic analysis techniques which can facilitate the search process at semantic level. We also study the problem that short queries are less informative and difficult to convey the user's intention into the search service system. We propose a generalized framework to address these issues. We conduct a case study of movie plot search in which a semantic analyzer seamlessly works with a user's intention detector. Our experimental results show the importance and effectiveness of intention detection and semantic analysis techniques.



# Acknowledgements

---

I wish to express my appreciation and gratitude to the thesis supervisors Dr. Jo Abrantes and Prof. Philip Ogunbona for their guidance and encouragement throughout the course of this research work. Dr. Jo Abrantes introduced me into the field of machine learning and offered me the chance to deepen my research. Prof. Philip Ogunbona gave me a lot of constructive suggestions and ideas. Especially they gave me as much freedom as I needed to get my research work done.

I gratefully thank my parents for their support. Without their understanding, encouragement and financial support, this particular goal would not have been achieved.

# Contents

---

<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>1 Introduction</b>	<b>7</b>
1.1 Motivation . . . . .	7
1.2 Contributions . . . . .	13
1.3 Organization of the thesis . . . . .	14
<b>2 Literature Review</b>	<b>16</b>
2.1 Introduction . . . . .	16
2.2 Overview of Search Service Models . . . . .	18
2.2.1 The Boolean Model . . . . .	19
2.2.2 The Vector Space Model . . . . .	19
2.2.3 Probabilistic Models . . . . .	22
2.2.4 The Link Analysis Models . . . . .	22
2.3 Semantic Analysis Techniques . . . . .	23

---

2.3.1	Explicit Semantic Analysis . . . . .	24
2.3.2	Latent Semantic Analysis . . . . .	26
2.4	Some Related Techniques In Search Services . . . . .	31
2.4.1	Relevance Feedback . . . . .	31
2.4.2	Filtering Technique . . . . .	32
2.4.3	Question-Answering System . . . . .	32
2.4.4	Text Classification . . . . .	33
2.5	Chapter Summary . . . . .	36
<b>3</b>	<b>Intention Detector Construction</b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	Determination of Type of User Intentions . . . . .	40
3.3	Detection of the State of the Specific User Intention . . . . .	41
3.4	The Simplified Intention Detector . . . . .	45
3.5	Selection Of Feature Words . . . . .	47
3.6	Chapter Summary . . . . .	49
<b>4</b>	<b>A Case Study Of Movie Plot Search</b>	<b>50</b>
4.1	Introduction . . . . .	50
4.2	Query Parser . . . . .	51
4.3	Movie Plots Tagger . . . . .	52
4.4	Movie Plot Index Builder . . . . .	52
4.4.1	Tokenization and Stopword Removal . . . . .	53
4.4.2	Stemming . . . . .	53

---

4.5	User Intention Detector . . . . .	54
4.6	Query Formulation . . . . .	57
4.7	Semantic Analysis . . . . .	58
4.8	Ranked Result Recommender . . . . .	61
4.9	Chapter Summary . . . . .	61
<b>5</b>	<b>Experimental Results</b>	<b>62</b>
5.1	A Visualization Experiment . . . . .	62
5.2	An Intelligent Search Service Demonstration . . . . .	65
5.3	An Evaluation Experiment . . . . .	73
5.4	Some Experiments About User Queries . . . . .	76
5.5	Chapter Summary . . . . .	81
<b>6</b>	<b>Conclusions and Future Work</b>	<b>82</b>
	<b>Bibliography</b>	<b>84</b>

# List of Tables

---

5.1	One Experiment for Two Small Document Collections . . . . .	63
5.2	The Term-document Matrix . . . . .	63
5.3	Top Ten Most Representative Words . . . . .	78
5.4	Find Most Related Words Via Jiang's Method . . . . .	80
5.5	Find Most Related Words Via Lin's Method . . . . .	81

# List of Figures

---

1.1	The Search Processing . . . . .	9
2.1	The Link Structure of World Wide Web . . . . .	23
2.2	The Process of Latent Semantic Analysis . . . . .	29
3.1	User Intention Detection Procedure . . . . .	39
3.2	Illustration of Determination of Type of User Intentions . . . . .	40
3.3	Illustration of First Order Hidden Markov Chain for Prediction . . . . .	42
3.4	Illustration of Independent Node Model for Prediction . . . . .	43
3.5	Illustration of Second Order Hidden Markov Chain for Prediction . . . . .	44
3.6	Illustration of the Simplified Intention Detector . . . . .	46
3.7	Illustration of Keywords Representation of the Simplified Intention Detector . . . . .	47
4.1	Architecture of Intelligent Search Service System . . . . .	51
4.2	Architecture of Search Service . . . . .	54
4.3	Illustration of the Simplified Intention Detector . . . . .	55

---

5.1	Visualization of Latent Semantic Analysis . . . . .	64
5.2	The Initial Search User Interface . . . . .	67
5.3	The Concept Search Results . . . . .	69
5.4	The Topic Search Results . . . . .	70
5.5	The Content Search Results . . . . .	71
5.6	The Content Search Results . . . . .	72
5.7	The Precision and Recall . . . . .	74
5.8	The Precision and Recall Results . . . . .	75