碩士學位 論文

통계패키지 R과 유사도 알고리즘을 활용한 추천시스템 구현

-생명보험 사례를 중심으로-

Implementation of the Recommendation System Using the Statistical Package R and Similarity Algorithm

-Focus on Life Insurance-

2014 年 12月

崇實大學校 情報科學大學院

소프트웨어工學科

李 承 珉

碩士學位 論文

통계패키지 R과 유사도 알고리즘을 활용한 추천시스템 구현

-생명보험 사례를 중심으로-

Implementation of the Recommendation System Using the Statistical Package R and Similarity Algorithm

-Focus on Life Insurance-

2014 年 12月

崇實大學校 情報科學大學院

소프트웨어工學科

李 承 珉

碩士學位 論文

통계패키지 R과 유사도 알고리즘을 활용한 추천시스템 구현

-생명보험 사례를 중심으로-

指導敎授 崔 龍 洛

이 論文을 碩士學位 論文으로 제출함

2014 年 12月

崇實大學校 情報科學大學院

소프트웨어工學科

李 承 珉

李 承 珉의 碩士 學位 論文을 認准함

審 査 委 員 長 印

審 査 委 員 印

審 査 委 員 印

2014 年 12月

崇實大學校 情報科學大學院

목 차

國文抄錄 ····· iv
英文抄錄 ····· v

제 1 장 서 론

1.1 연구 배경 및 목적 ····· 1
1.2 연구 내용 및 구성 ····· 2

제 2 장 관련 연구

2.1 공개용 통계패키지 R ····· 3
2.2 알고리즘 개요 ····· 8
2.3 생명보험 소개 및 업무 서비스 ····· 14

제 3 장 R을 활용한 추천시스템 구현

3.1 R과 JAVA AWT를 활용한 데이터 분석 ····· 16
3.2 R과 유사도 알고리즘을 활용한 추천시스템 구현 ····· 29
3.3 협업적 필터링과 내용기반 필터링 추천시스템의 비교 ····· 34

제 4 장 결론 및 향후 과제

참고문헌 ····· 39

표 목 차

[표 2-1] R의 역사 ····· 3
[표 3-1] 2014년 연령별 실적 가입설계 데이터 ····· 24
[표 3-2] R의 자료구조 ····· 26
[표 3-3] Rserve API ····· 26
[표 3-4] REXP API ····· 29
[표 3-5] RList API ····· 27
[표 3-6] 코사인 유사도를 행렬로 정규화한 데이터 ····· 30
[표 3-7] 연령대 그룹 사이간 코사인 유사도를 대입한 결과 ····· 31
[표 3-8] 유사도 알고리즘 구현 소스 및 설명 ····· 31

그 림 목 차

[그림 2-1] R의 콘솔화면 ····· 5
[그림 2-2] R의 작동방식 ····· 6
[그림 2-3] 협업적 필터링과 내용기반 필터링 ····· 9
[그림 2-4] 추천 시스템의 개념 ····· 11
[그림 3-1] R 홈페이지 사이트 ····· 16
[그림 3-2] R 설치파일 다운로드 ····· 17
[그림 3-3] R 설치 화면 ····· 18
[그림 3-4] 라이센스 동의화면 ····· 18
[그림 3-5] 설치 디렉토리 선택화면 ····· 19
[그림 3-6] 구성요소 선택화면 ····· 19
[그림 3-7] R 3.3.1 기본 실행화면 ····· 20
[그림 3-8] R 3.3.1 Rserve 라이브러리 설치 및 서버 실행화면 ····· 21
[그림 3-9] 이클립스에서 JAVA R 연동 실행화면 ····· 22
[그림 3-10] R 2.4.0에서 Rserve 서버 설치화면 ····· 23
[그림 3-11] R 2.4.0 Rserve 서버 실행화면 ····· 23
[그림 3-12] R 2.4.0 JAVA AWT로 구현한 표 차트 생성 ····· 26
[그림 3-13] R 접속 테스트 ····· 24
[그림 3-14] R Chart를 생성하는 소스 ····· 29
[그림 3-15] 순서도 및 추천 시스템 소스 ····· 32
[그림 3-16] 고객정보 입력화면 ····· 33
[그림 3-17] 결과 값 내용확인 ····· 33
[그림 3-18] R 패키지와 코사인 유사도 기반 추천시스템 ····· 34

文抄錄

통계패키지 R과 유사도 알고리즘을 활용한 추천시스템 구현

-생명보험 사례를 중심으로-

소프트웨어工學科　李承珉

指 導 敎 授　崔龍洛

오늘날 급속도로 사회의 변화가 이루어짐에 따라 기업에서는 미래를 예측하여 상품에 대한 고객의 니즈를 충족시켜야 하는 필요성이 제기되고 있으며, 앞으로 고객 개개인에 대한 맞춤 비중은 더욱더 커질 것으로 예상된다.

현재 보험 산업은 소비자 변화, 보험시장 구조조정 본격화, 경쟁영역의 확대 등으로 급격한 성장을 이루었으며, 이에 대한 신속한 대응이 적절하게 이루어져야만 수익성을 보장 받을 수 있게 되었다. 상품 경쟁격화로 인해 소비자를 위한 복합형 상품이 등장하고, 이는 패턴이 다양한 고객 수준과 복합형 상품 처리능력에 대한 경쟁력이 결정 요소로 등장하는 계기로 작용되고 있다.

본 논문에서는 공개용 통계패키지 R과 유사도 알고리즘을 활용하여 연령대별 그룹을 나누어 입력된 고객정보와 가입유형이 유사한 그룹을 찾아 실적이 많은 상품을 추천 할 수 있는 시스템을 구현 하고자 한다.

ABSTRACT

Implementation of the Recommendation System Using the Statistical Package R and Similarity Algorithm

-Focus on Life Insurance-

LEE, SEUNG MIN

Department of Software Engineering
Graduate School
Soongsil University

As the society rapidly changes with the development, the significance of predicting consumer's needs as well as fulfilling the needs has been an issue for a company. Due to that reason, we expect that customized service on each customer will be larger from now on. Current insurance industry made a success after consumer trend change, insurance market restructuring, expanding work area including competitor, and so on and therefore a rapid reaction regarding to those factors should follow in order to make certain profit. Due to product's tight competition, combined products appeared in the market. As a result, the combination of diversity of consumer's range and the each product's competitiveness become a final decision factor for a customer.

In this article, as using the public Statistical Package R and Similarity Algorithm, the system will categorize consumers into age groups. And the program will sort out people who have similar taste based on the entered customers' information, And finally the system will be able to suggest the best-selling products of all time.

제 1 장 서 론

1.1 연구의 배경 및 목적

최근, 포화상태에 진입한 생명보험 시장에서 유사한 상품구조로는 더 이상 수익을 담보할 수 없는 상황에 이르렀다. 단순히 경쟁보험사들과 비교하여 보험 상품의 구색만 갖추고 이벤트를 강화한다거나 판매채널을 추가하여 고객을 모집하는 기존 방식 틀에서 벗어나야 한다.

과거에는 고객들이 보험 상품에 대한 비교 및 합리적인 구매 의사 결정이 용이하지 않았으나, 고객들의 인터넷 거래 선호 현상 등으로 금융서비스 이용에 라이프스타일이 변화되었고, 개인별로 처해진 자신의 상황들을 신속하게 인지해주고 즉각적으로 반응해주는 서비스 형태들이 각광을 받고 있다.

생명보험의 마케팅 전략은 기존의 계약중심에서 '고객중심의 가치성장'으로 변화 되었으며, 보험 상품을 가입하려는 소비자들은 스마트폰 및 SNS(Social Networking Service) 등을 통한 실시간 정보 공유를 하면서, 커뮤니티 단위의 집단적인 참여로 니즈 반영이 활발해져 보험 상품에 대한 충분한 커뮤니케이션과 가격비교도 가능해 졌다.

이러한 데이터를 활용하여 차세대 시스템 개선에 실시간으로 적용하려는 기업이 증가함에 따라 고객 맞춤형 서비스를 제공하고 있으며, 일회성으로 사용되고 무의미하게 버려지는 많은 데이터 속에서 연관성을 찾아 의미 있는 정보를 추출하고 결과를 분석하려는 노력이 필요하다.

본 논문에서 구현 시스템에 활용할 통계패키지 R은 1993년 뉴질랜드 오클랜드대학의 통계학과 교수 2명(Roos Ihaka, Robert Gentleman)에 의하여 개발 되었으며, 1976년 Bell Lab의 John Chambers, Rick Becker,

Allan Wilks에 의하여 개발한 S Language 기반에 통계분석 모듈과 함께 다양한 그래픽 도구를 사용할 수 있다.[1] 특히, 구글과 페이스북은 R을 주된 분석 플랫폼으로 사용하고 있으며, 일련의 데이터 처리 및 분석 작업을 대화 형으로 처리할 수 있고, 뛰어난 그래픽 기능을 가지고 있어 그래프를 활용한 자료 분석 작업도 매우 쉽게 할 수 있는 장점을 가지고 있다.

1.2 연구 내용 및 구성

본 논문은 1장에서 연구 배경과 목적 그리고 연구 방법을 기술하였다. 2장에서는 공개용 통계패키지 R에 대한 설명 및 생명보험 시스템을 소개한다. 3장에서는 R과 유사도 알고리즘을 활용하여 추천시스템을 구현해 보고, 추출된 문제점들에 대해 검토 후 4장에서는 결론과 향후 과제를 제시하였다.

1) 이동우, 「R, 그리고 빅데이터」, 제 30회 Open Technet 빅데이터 오픈소스 플랫폼 기술세미나, 2012, 14p.

제 2 장 관 련 연 구

2.1 공개용 통계패키지 R

2.1.1 R 소개

통계 패키지란 복잡한 통계적인 절차와 연산처리를 대신 해주는 자료 분석 프로그램이다. 그 중에서 공개용 통계패키지 R은 수치 연산, 데이터의 관리 그리고 시각화를 통합적으로 지원하는 소프트웨어로써 다양한 그래픽 도구로 이루어져 있으며, 유료 통계 패키지인 프로시저(Procedure)중심의 SAS(Statistical Analysis System), 메뉴 중심적인 SPSS(Statistical Package for the Social Science) 와는 다르게 R은 Interpreter Language 분석 시스템에 기반을 두고 있다.

[표 2-1] R의 역사

년도 / 언어	S Language	S-PLUS	Package R
1976	John Chamber		
1980	Ver1, Fortran-based		
1988	Ver2, UNIX	StatSci	
1993	Ver3, C-base	With MathSof E-license	Ross Ihaka Roert Gentleman
1997.4.1			Mailing list
1997.12.5			CRAN
2001	Ver4, Java Interface	Insightful	GNU Project
2008		OS.V.7 Big data 07 V.8.R package	Ver1

R은 1976년 미국 AT&T 벨 연구소의 존 챔버스, 릭 베커, 앨런 윌크

스에 의해 개발된 S언어를 기반으로, 현재 사용되는 R은 1993년 뉴질랜드 오클랜드 대학의 로스 이하카(Roos Ihaka)와 로버트 젠틀맨(Robert Gentleman)에 의해 시작되어 GPL(General Public License)에서 배포되는 GNU(General Public License) S 라고도 한다. 개발자 이름의 이니셜 알파벳을 따 R로 명명 되었다.[2)]

2.1.2 R 특징

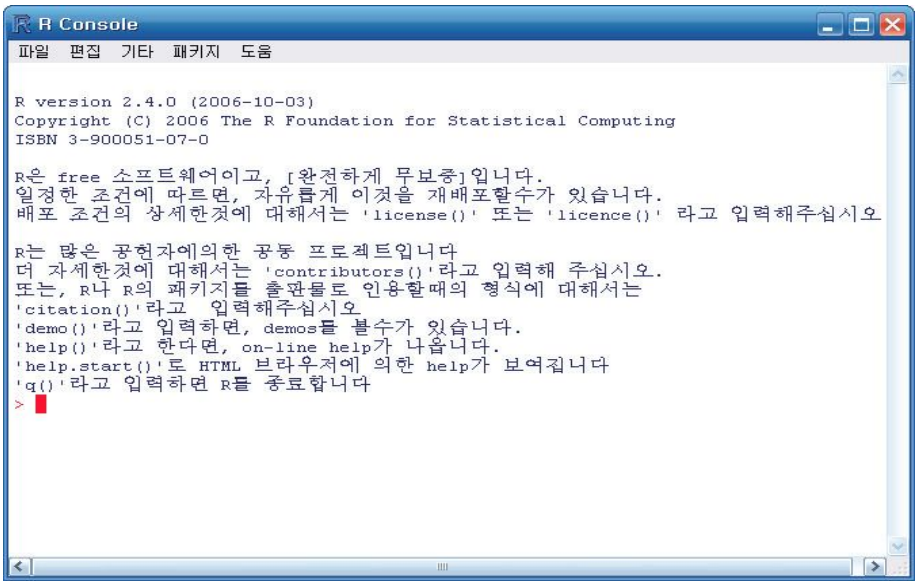
통계분석 소프트웨어로 R을 선택한 이유 및 최대 장점은 오픈소스이기 때문에 어떠한 행위에 제한을 받지 않을 뿐더러 새로운 기능 추가가 더 빠르고 가벼우면서도 다양한 기능을 갖춘 R은 자료 분석, 통계 모델링, 데이터 처리, 그래프의 결과물 출력 등으로 여러 가지 목적으로 사용된다. R은 시스템 통합의 용이성이 있으며, 그 통합 사례로는 Revolution Analytics의 Revolution R, IBM의 Netteza Appliance DB, EMC의 Greenplum Appliance DB를 예를 들 수 있다. 통계분석에 최적화된 자료구조로 Matrix, Vector 등을 Data Objects로 사용할 수 있으며 행렬/벡터 데이터 타입 지원과 행렬 연산 지원으로 복잡한 구조의 반복문을 제거하고 코드를 이해하기가 쉽게 구현 되었다. 현재는 버전이 증가함에 따라서 기하급수적으로 R 패키지 수가 증가하고 있는데 R 패키지를 실시간으로 다운로드 받아 바로 사용할 수 있는 부분이 R언어가 데이터 분석 분야에서 가장 강력한 플랫폼이 된 주요 이유이다.

R의 단점으로는 메모리 한계가 이슈인데 데이터 처리작업 방식이 모든 데이터를 메모리에 로딩 후에 진행되기 때문이다. 데이터를 10GB 이상 처리 가능하나 너무 느리다는 단점이 있다. 불필요한 데이터 저장으

2) 이동우, 「R, 그리고 빅데이터」, Open Technet 빅데이터 오픈소스 플랫폼 기술세미나, 2012

로 인한 메모리 부족 현상이 있으며, 32비트에서 표현 가능한 숫자만이 사용되고 있다.

R은 구성은 일반적인 컴퓨터 프로그래밍 언어로 되어 있어, 일정 부분 프로그래밍 능력을 갖춰야 하지만 다른 개발언어와 달리 대부분 인식성이 높은 이름의 함수로 이루어져 있어 프로그래밍 기초가 약한 사람도 2일 정도 기본 교육을 받으면 기본적인 사용은 가능하다. 주로 통계학 전공자들이 사용하지만 비전공자들도 쉽게 이용할 수 있도록 되어 있다.

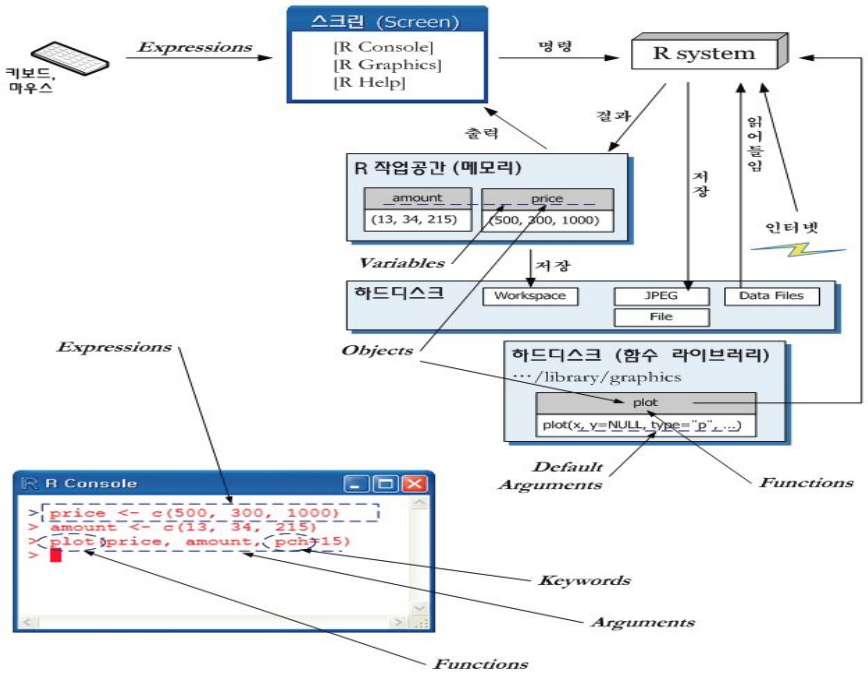


[그림 2-1] R의 콘솔 화면

R을 실행시키면 처음 접하게 되는 화면은 [그림 2-2] 과 같은 「R Console」형태이다. 위와 같이 명령을 기다리는 프롬프트를 볼 수 있는데, R언어는 객체 지향 프로그램 언어로써 데이터의 선언 및 처리작업 진행을 UNIX 시스템과 같이 대화형 커맨드로 처리할 수 있다.

2.1.3 R 작동방식

R을 이용해서 통계분석을 수행한다는 것은 업무 통계분석 절차를 R 함수를 활용하여 올바르게 사용한다는 것을 의미한다. 데이터의 입력 및 결과 출력과 관련된 R의 명령어와 함수를 자유자재로 사용하기 위해서는 R에 대한 기본적인 처리 절차를 설명하였다.[3)]



[그림 2-2] R의 작동방식

[그림 2-2]은 키보드 또는 마우스로 입력한 각종 표현식이 R 시스템에 요청하여 그래프 형식 또는 텍스트의 결과물을 스크린으로 보게 되는 R

3) 조완일, 「R의 설치 및 기본 사용법」, 주식회사 센소메트릭스, 2006

의 작동방식을 도식화 한 것이다. 표현식은 R 시스템이 이해할 수 있는 명령어를 말한다. 이 명령어 문장에는 처리 결과를 저장할 위치가 지정될 수도 있고 데이터가 포함될 수도 있다. 어떻게 요청하라는 지시사항은 당연히 포함되어 있을 것이다. 사람과의 대화와 마찬가지로 이러한 표현식(Expressions)은 정해진 문법에 따라 작성되어야 한다. 변수(Variables)는 데이터를 저장할 수 있는 저장 공간을 말한다. 개체란 데이터(Data, Dataset), 변수(Variables), 결과(Results), 함수(Functions)처럼 현재 사용 중인 컴퓨터 메모리(작업 공간)에 특정 이름으로 저장되는 것을 의미한다. R을 사용하다 보면 함수(Function)을 이용하는데 q() 라는 명령어 함수는 R 시스템을 종료할 때 사용하는 함수인데 괄호 안에 아무런 인자(Arguments)가 없이 사용되는 함수가 있는가 하면 plot(price, amount, pch=15)처럼 반드시 인자가 입력해야 하는 함수도 있다. plot 함수를 사용할 때 type="p" 라는 인자에 값을 입력하지 않아도 Default 값으로 포인트 타입의 그래프를 출력해 주는데 이는 plot 함수가 기본인자(Default Arguments)값을 가지고 있기 때문이다. args(plot default) 명령어를 사용하면 plot 함수의 기본인자를 확인할 수 있다. 기본인자는 함수 정의에 포함되어 있다.[4)]

2.1.4 R 사용 환경

R은 UNIX, Mac, Windows 등의 운영체제에서 사용 가능하며, Java, C, Fortran 등 프로그래밍 언어와의 인터페이스를 제공하므로 확장에 용이하다. 또한 오라클과 같은 DBMS(DataBase Management System) 와도 연동이 가능하고, Microsoft 의 Excel 이나 Access 같은 소형 데이터베

4) 조완일, 「R의 설치 및 기본 사용법」, 주식회사 센소메트릭스, 2006, 38p.

이스와도 연동이 가능하므로 사용자는 데이터를 쉽게 불러와 통계 분석을 할 수 있다.

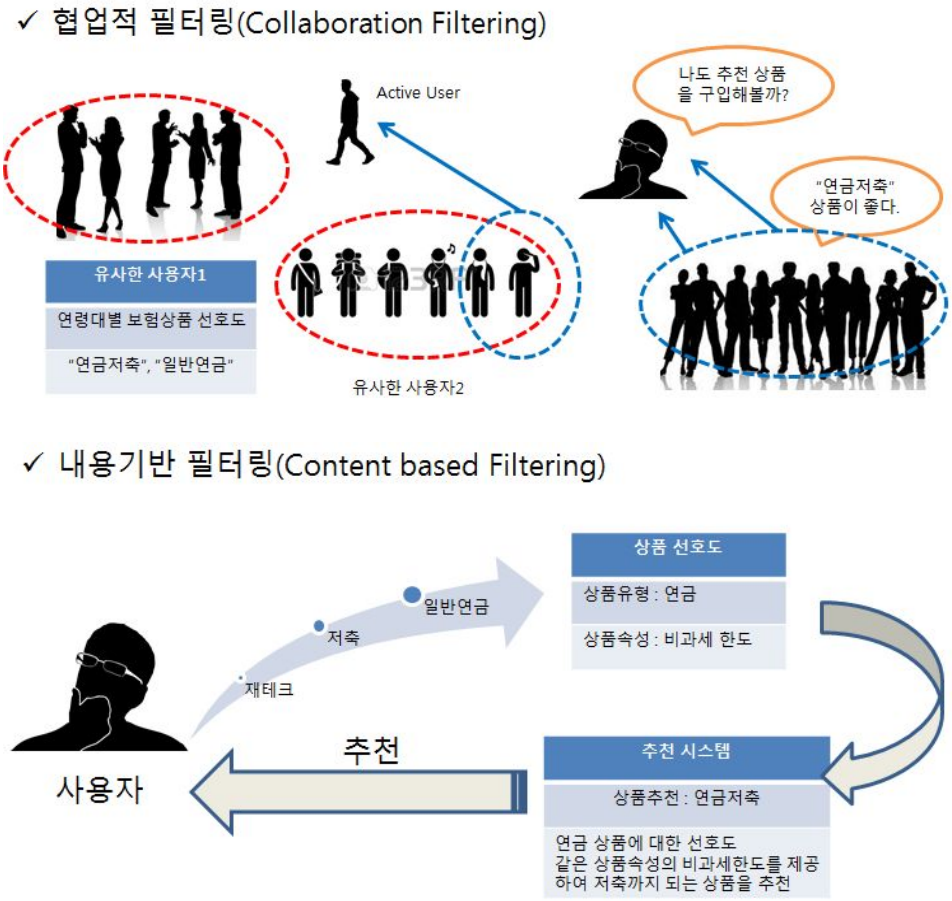
R은 CRAN Site(http://cran.nexr.com/) 에서 다운로드 받아 설치 할 수 있으며, 이동식 하드 디스크, USB(Universal Serial Bus) 메모리 등에도 설치가 가능하기 때문에, 로컬 하드디스크에 설치한 후 R 폴더 전체를 USB(Universal Serial Bus) 이동식 저장장치에 복사하여도 R 설치 폴더의 bin 하위폴더에서 실행시키면 사용에 지장이 없다.

2.2 알고리즘 개요

알고리즘이란 어떠한 문제를 해결하기 위해 컴퓨터 전산 프로그래밍이 수행해야 할 절차들을 나타낸 것이다. 순서에 따라 기계적으로 처리하려면 목적에 대한 결과를 도출할 수 있을 때 그 순서의 목적을 알고리즘이라고 한다.

2.2.1 추천 알고리즘에 관한 연구

일반적으로 추천 알고리즘은 내용기반 필터링(Content Based Filtering), 협업적 필터링(Collaborative Filtering)으로 구분할 수 있다.



[그림 2-3] 협업적 필터링과 내용기반 필터링

[그림 2-3]에 협업적 필터링은 이웃(Neighbor) 사용자들의 선호 프로파일 설정하여 그들이 선호하는 상품을 그 사용자에게 추천하는 방식이다. 이러한 설정을 참고하여 각각의 사용자와 성향이 유사한 사용자를 찾아낸 후, 각각의 가중치 평균으로 사용자가 접하지 못한 아이템 또는 상품에 대한 선호도를 예측하게 되는 것이다. 각각의 유사도에 따라 이웃 사용자가 다르게 형성되고 이에 따라서 선호도 점수들이 상이하게 나타날

수 있으므로, 이웃형성 과정은 협업적 필터링에서 필수 요소라 할 수 있다.[5)]

내용기반 필터링은 사용자간의 선호도 정보를 사용하는 것이 아니라 사용자 별 상품의 속성 또는 특정 콘텐츠를 기반으로 한다. 내용기반 필터링 기법은 정보의 검색 또는 정보의 추출 분야에서 발전된 것으로써, 관련 콘텐츠 또는 관련 상품의 추천을 위해 사용자가 요구하는 정보간의 유사도를 계산한 결과를 순위화하여 표현한다. 콘텐츠의 내용을 기반으로 추천하는 방법으로써, 이를 구현하기 위해서는 적합성 피드백, 가중치 기법, 확률검색 모형 등이 활용된다.

2.2.2 추천 시스템에 관한 연구

추천 시스템이란 인터넷 사용자들을 대상으로 고객에게 상품을 추천하고, 고객들이 상품 구매 선택을 돕기 위해 정보를 제공하는 시스템으로 정의할 수 있다. 이러한 추천 시스템은 다음과 같이 구매 유도를 촉진을 위해 3가지 방식으로 확인할 수 있다.[6)]

(1) 구매 유도

단순히 사이트를 구경하러 온 사용자에게 구매를 유도하기 위해 적절한 상품 정보를 노출시킬 수 있다.

(2) 교차 판매

추천 시스템은 새로운 구매를 유도하기 위해 구매예정 고객에게 부가적으로 다른 적절한 상품을 추천할 수 있다.

---

5) 김용수, 「개인화 서비스를 위한 추천시스템의 연구동향」, 대한산업공학회, 2012

6) 김병국, 「유전자 알고리즘을 활용한 개인화된 상품추천시스템 개발」, 동국대학교, 2004, 6p

(3) 로열티 증대

추천 시스템은 고객에게 고객의 로열티를 증대하기 위한 가치 있는 정보들을 함께 제공할 수 있다.

[그림 2-4] 추천 시스템의 개념

[그림 2-4]에 추천 시스템은 고객 맞춤형 서비스로 고객의 연령대, 구매 상품에 대한 선호도 등을 파악하여야 하며, 고객 개개인의 동적인 행위의 대한 분석을 토대로 다양한 관점에서의 선호도를 추출 및 활용함으로써, 각 개인별로 다른 상품 및 콘텐츠를 제공하는 것이다. 첫 번째로 고객이 어떠한 상품을 선호하는지 도출해야 한다. 특정한 아이템에 대한 선호도뿐만 아니라, 관련된 콘텐츠 속성에 대한 선호도를 모두 도출하는 것을 의미한다.

예를 들어 특정 사용자의 보험 상품에 대한 선호도라고 하면, 보험 대상자의 인적정보, 면담정보, 접촉정보, 주소, 수익자정보 등을 가고 있는 연령대 그룹을 만들어 그들이 가입한 상품정보의 실적을 확인하여, 각

사용자의 유사한 선호패턴을 갖는 집단인 그룹을 결정한다. 이는 특정 연령대 사용자에게 적합한 상품을 추천하고자 할 때, 기존에 그 사용자와 유사한 성향의 사용자들이 선호했던 상품을 추천하기 위함이다.

추천 시점으로써는 사용자가 상품 또는 서비스를 어떤 상황(Context)에서 구매하는지 파악함으로써 추천 시점을 결정할 수 있다.[7)]

2.2.3 코사인 유사도

두 벡터의 유사도를 구하는 방법으로써 유사도를 구할 때 두 벡터 사이의 각을 코사인(Cosine)값을 구해서 유사도로 취급하기 때문에 코사인 유사도(Cosine Similarity)라고 한다. 코사인 유사도의 공식은 다음과 같다.

$$similarity = \cos(\Theta) = \frac{A \cdot B}{\|A\|\|B\|}) = \frac{\sum_{i=1}^{n} A_i \times B^i}{\sqrt{\sum_{i=1}^{n}(A_i^2)} \times \sqrt{\sum_{i=1}^{n}(B_i^2)}}$$

분자 부분과 분모 부분을 나눠서 설명하면,

분자는 아래와 같은 두 벡터가 있다면, 각 원소들 순서대로 짝을 맞춰서 곱한 다음에 결과들을 다 더하면 된다.

A = {1, 2, 3, 4, 5}, B = {6, 7, 8, 9, 10}

위의 벡터에 대한 데이터를 각각 곱셈을 한다.

1 * 6 = 6

2 * 7 = 14

7) 김용수, 「개인화 서비스를 위한 추천시스템의 연구동향」, 대한산업공학회, 2012

3 * 8 = 24

4 * 9 = 36

5 * 10 = 50

곱한 것을 다 더하게 되면, 130이란 결과 값이 나오게 된다.

6 + 14 + 24 + 36 + 50 = 130

분모 (두 벡터의 크기를 곱한다.)

||A|| 는 A벡터의 크기를 말하고, ||B|| 는 B벡터의 크기를 말한다.

분모는 두 벡터의 크기를 구해서 곱하면 된다.

삼각형의 빗변의 길이 구하기를 기억한다면,

$C = \sqrt{A^2 + B^2}$

A 벡터의 크기를 구한다.

$\sqrt{1^2+2^2+3^2+4^2+5^2} = 7.416$

B 벡터의 크기를 구한다.

$\sqrt{6^2+7^2+8^2+9^2+10^2} = 18.1659$

A 벡터의 값과 B 벡터의 값을 곱한다.

7.417 * 18.1659 = 134.7219

마무리로 분자를 분모로 나눈다.

$\frac{130}{134} = 0.9649$

위에 나온 결과 값이 코사인 유도 값이다.

2.3 생명보험 소개 및 업무 서비스

문명이 발달할수록 인간사회에는 갖가지 위험이 생겨나 오늘날 우리들은 무수히 많은 위험에 직면하면서 살아가고 있다. 각종 재해로부터 교통사고, 질병, 상해, 사망과 같은 불의의 사고에 이르기까지 그 경우의 수는 너무나 많다. 이로 인하여 가족들의 생계를 책임지고 있는 가장들의 경제생활도 항상 불안할 수밖에 없는 상태이다. 따라서 언제 어느 때 이러한 사고나 재난에 직면하게 되더라도 안정적인 경제생활을 유지할 수 있도록 이에 대한 충분한 준비를 사전에 해두는 것이 필요하다. 혼자서 이러한 위험에 대비하는 것은 한계가 있을 수밖에 없으므로 자연히 인류는 여럿이 힘을 합쳐서 위험에 대비하는 방법을 찾게 되었고, 그 결과 생겨난 것이 보험(Insurance, Versicherung, Assurance)이라는 금융제도이다.

생명보험은 이와 같이 사람의 생사(生死)에 관한 사고로 인해 초래되는 경제적 손실을 보호하기 위해 성립된 제도인데, 많은 사람들이 모여 합리적으로 계산된 소액의 보험료를 각출하여 공동으로 재산을 준비한 후 불의의 사고를 당한 사람에게 약정된 보험금을 지급해 줌으로써 안정적인 가족생활을 유지할 수 있도록 도와주는 경제준비의 사회적 형태라 할 수 있다.

생명보험 업무는 보험 적용이 되는 피보험자를 대상으로 사전에 보험가입설계를 할 수 있으며 고객정보 및 납입기간, 만기일자로 예상되는 해지환급금을 고객이 확인하면, 신계약, 청약을 신청하게 되는데 우선적으로 고객등록이 이루어져야 하고, 전산 자동심사와 보험에 대한 초회보

험료 입금이 완료되면, 보험 심사부서의 심사를 거쳐, 계약이 성립 후 계약보전에서 관리가 된다. 계약에 따른 부활 및 보험만기, 보험납기에 관련된 입금, 지급, 수수료가 발생되는 수익구조이다.

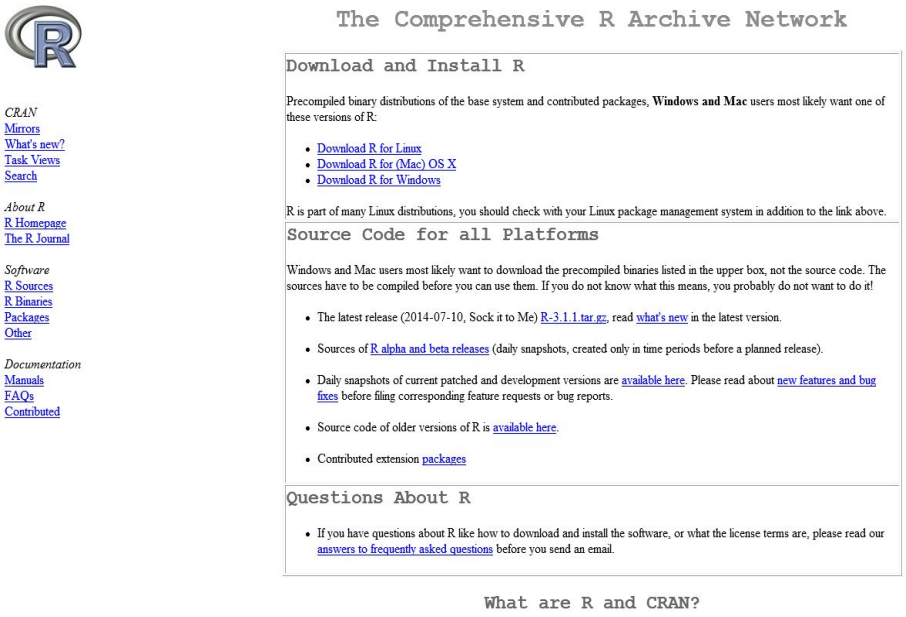
생명보험 시스템은 수많은 하위 시스템을 보유하고 있으며 각각의 업무 시스템마다 운영 환경이 다르게 적용되어 있다. 시스템을 구축했던 시기가 다를 수도 있지만, 효율적인 서비스를 제공하기 위해서 특화된 시스템으로 구축 되었다.

일반적으로 생명보험 시스템은 대외계, 채널계, 계정계 시스템으로 나누어져 있으며, 외부에서 고객, 사용자의 요청으로 들어온 데이터를 대외계 시스템에서 업무 데이터, 시스템 데이터를 각각 분리하고, 채널계 시스템에서 계정계 시스템에 필요한 데이터를 조립한 후 계정계 시스템 서비스 모듈에서 처리 후 응답 데이터를 고객화면에 재전송하는 방식을 가지고 있다. 이러한 실 계약이 성립 된 데이터를 가지고 적재 후 활용하여 영업 마케팅 지원·관리를 위해 사용되고 있다.

제 3 장 R을 활용한 추천시스템 구현

3.1 R과 JAVA AWT를 활용한 데이터 분석

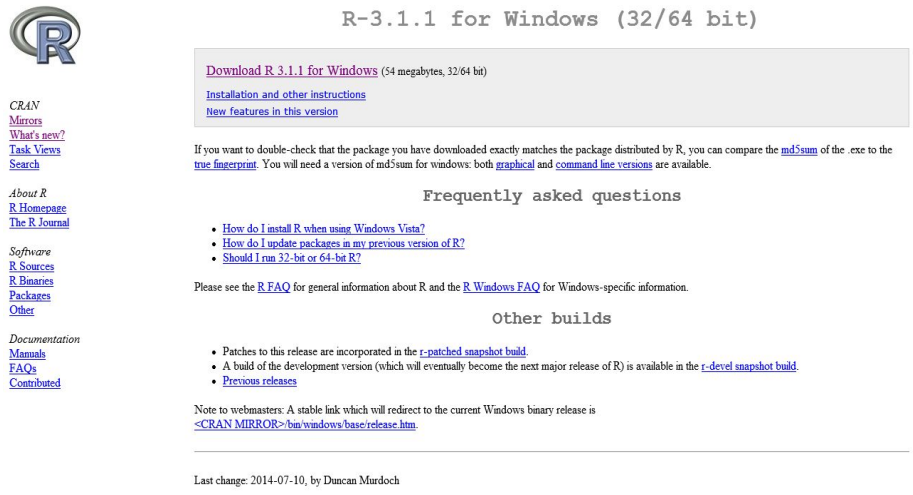
R을 활용한 데이터 분석을 사용하기 위해 다음과 같이 패키지 설치를 진행하였다. CRAN 홈페이지 사이트 “http://cran.r-project.org/”에 방문하면 컴퓨터 환경에 맞게 종류별로 제공하고 있다.



[그림 3-1] R 홈페이지 사이트[8)]

본 논문에서는 개인용 컴퓨터의 환경이 윈도우이기 때문에 [그림 3-1]에서 “Download R for Windows‘를 선택하였다.

8) <출처> http://cran.r-project.org/

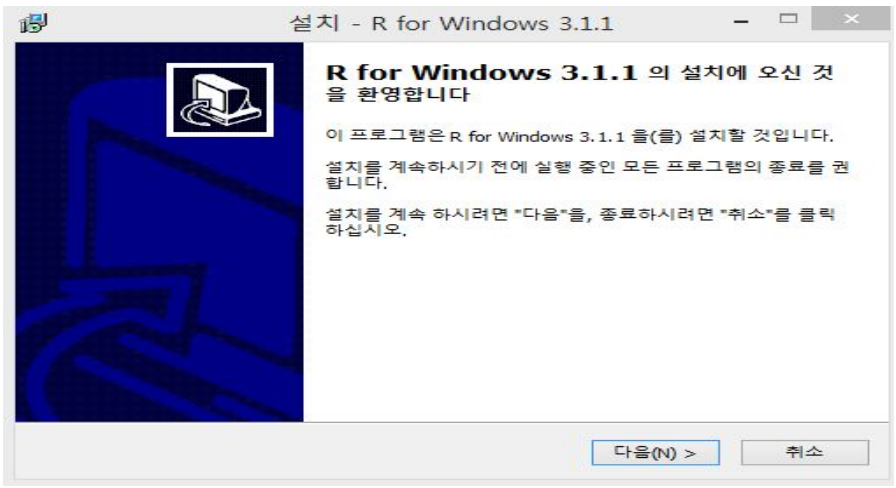


[그림 3-2] R 설치파일 다운로드[9)]

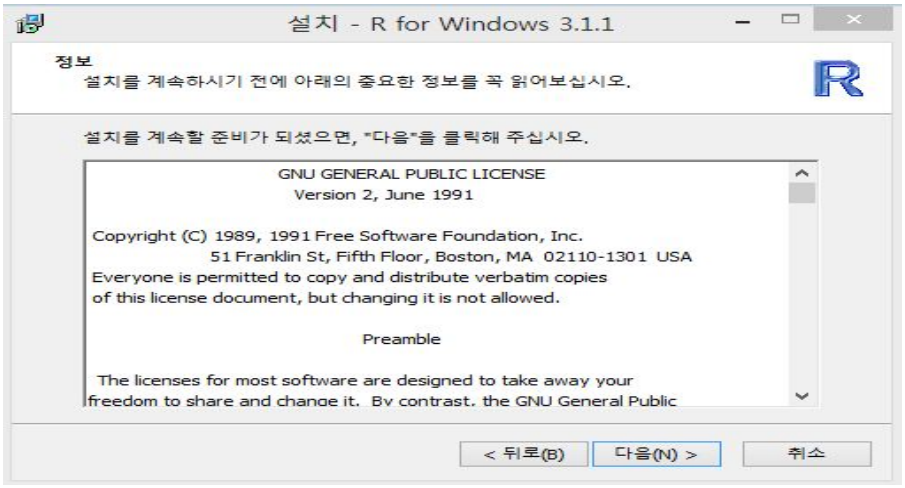
[그림3-2]에서 제공되고 있는 파일은 R 3.1.1 최신버전으로 컴퓨터 사양에 따른 32bit/64bit를 선택하여 설치할 수 있고, 32bit/64bit의 차이점은 컴퓨터 CPU하드웨어 처리작업에 따른 선택으로 윈도우 OS(Operating System) 제어판에서 확인할 수 있다. 화면 상단에 위치한 “Download R 3.1.1 for Windows″ 링크를 클릭하여 설치 파일을 다운로드 하였다.

R 홈페이지에는 각 나라별로 Mirors 사이트를 제공하기 때문에 본인이 속한 나라를 선택하여 다운로드를 빠르게 받을 수 있으며, 년도별로 등록된 R 패키지에 대한 설명 및 매뉴얼을 제공하고 있다.

9) <출처> http://cran.r-project.org/

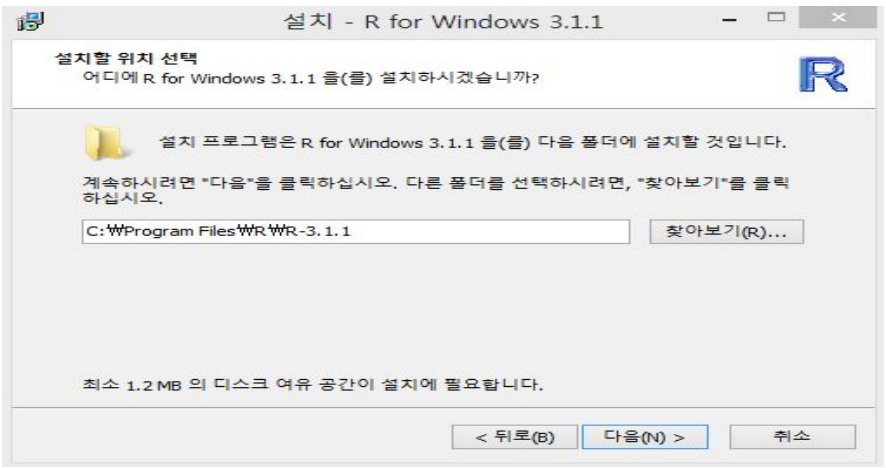


[그림 3-3] R 설치 화면



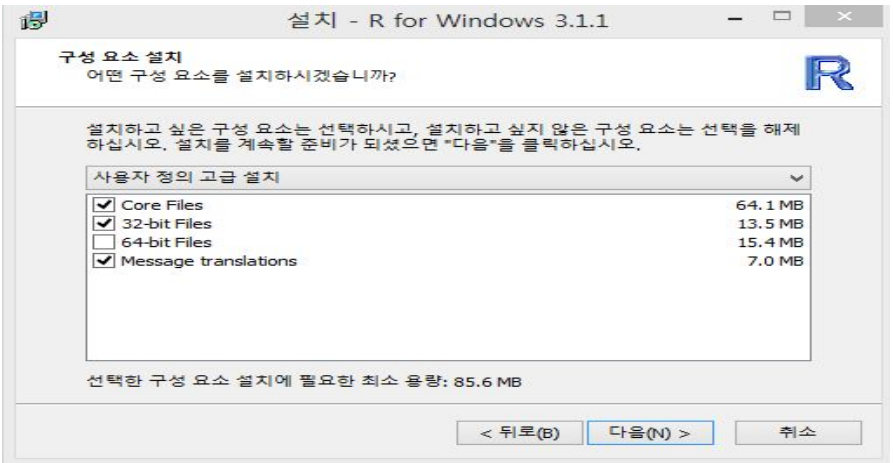
[그림 3-4] 라이센스 동의화면

[그림 3-4] GNU GENERAL PUBLIC LICENSE에 관련된 정책 안내 화면을 확인할 수 있으며, 사용자들이 자유롭게 배포가 가능하고, 공유할 수 있는 사항이 명시되어 있다.



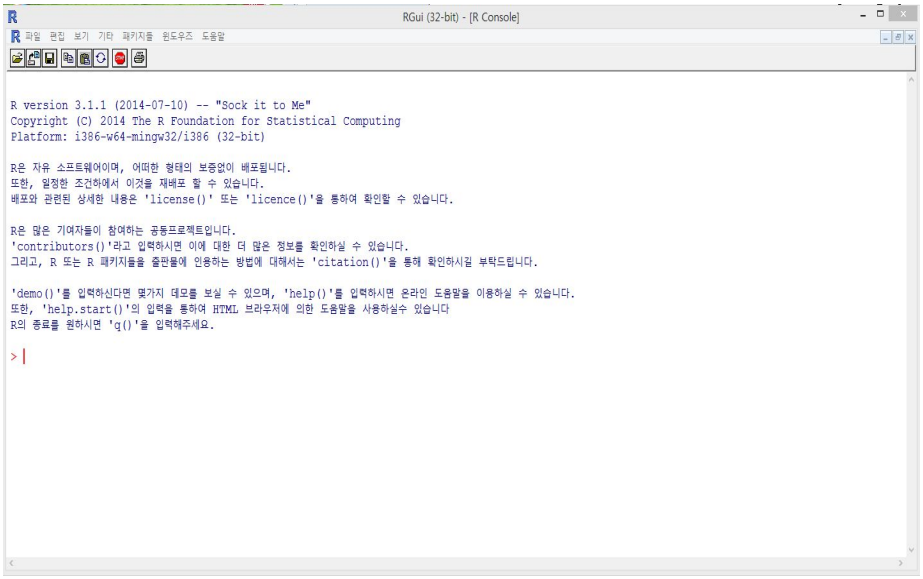
[그림 3-5] 설치 디렉토리 선택 화면

찾아보기 버튼을 클릭한 후 기본설치 폴더 외에 다른 폴더를 선택하거나 다른 드라이브를 선택할 수 있다.



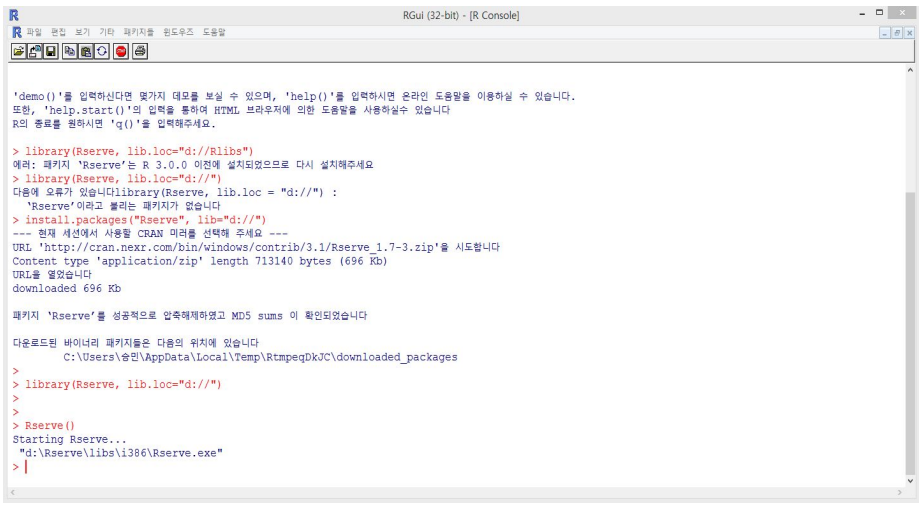
[그림 3-6] 구성요소 선택화면

[그림 3-6] 설치할 구성요소에서 설치하는 OS(Operating System)의 따른 32bit/64bit 파일이 맞는 지 확인하였다.



[그림 3-7] R 3.3.1 기본 실행화면

정상적으로 설치가 완료된 후 R을 실행(설치한 버전에 따라 R옆에 R 3.1.1 실행파일 이름에 표시됨)하면 [그림 3-7] 같은 R Console창을 확인할 수 있다. 다음으로는 유사도 알고리즘을 구현하기 위해서 R 패키지와 연동하여 자바로 구현하였으며, 자바와 연동하려면 R 패키지 API를 사용하면 된다. Rserve 관련 설치법은 이용필, 숭실대학원, 「통계패키지 R과 JSP를 활용한 데이터 분석 시스템 구현」, 2014 를 참고하여 설치를 하였다. Rserve 서버를 구동하기 위해서는 R 콘솔화면에서 Rserve 패키지를 설치 후 서버를 실행하였다.



[그림 3-8] R 3.3.1 Rserve 라이브러리 설치 서버 실행화면

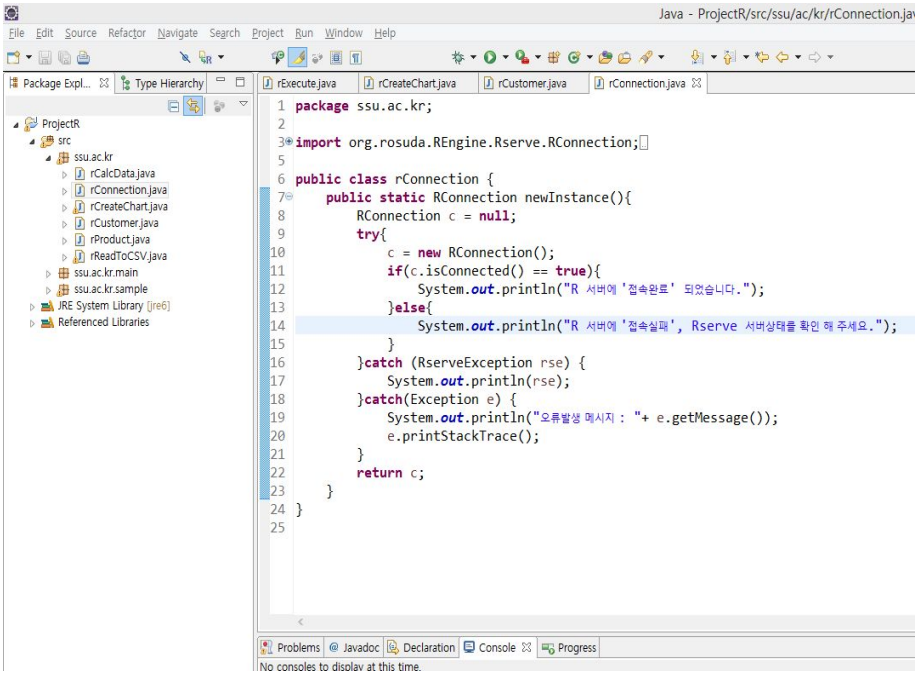
install.packages(“Rserve", lib="d://Rlibs")

위와 같은 R 명령어를 사용하여 라이브러리 폴더를 지정하였다.

library(Rserve, lib.loc="d://Rlibs")

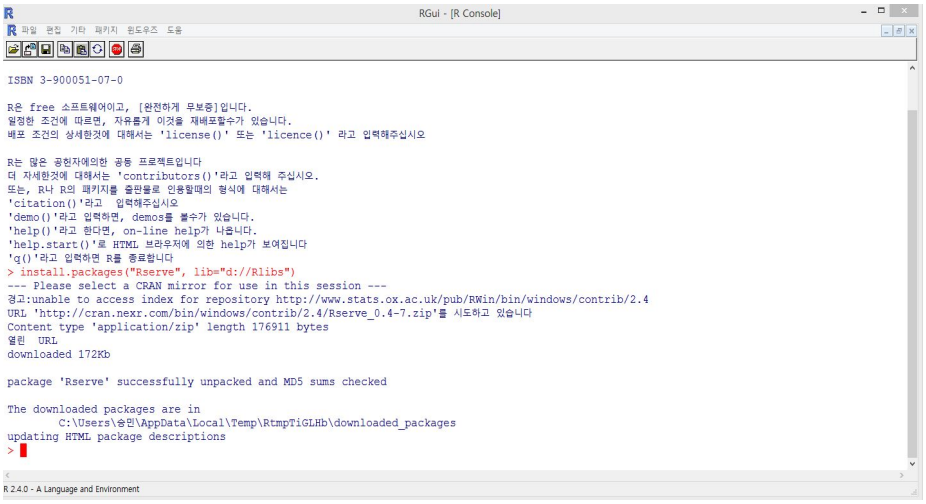
Rserve()를 실행시키면 서버가 구동되고 서버작업이 완료 후 이클립스(Eclipse)에서 Rserve 라이브러리 JRclient-RF503.jar를 Java Build Path에 포함하여 실행하였다.[10)]

10) 관련 jar 설치파일 http://rosuda.org



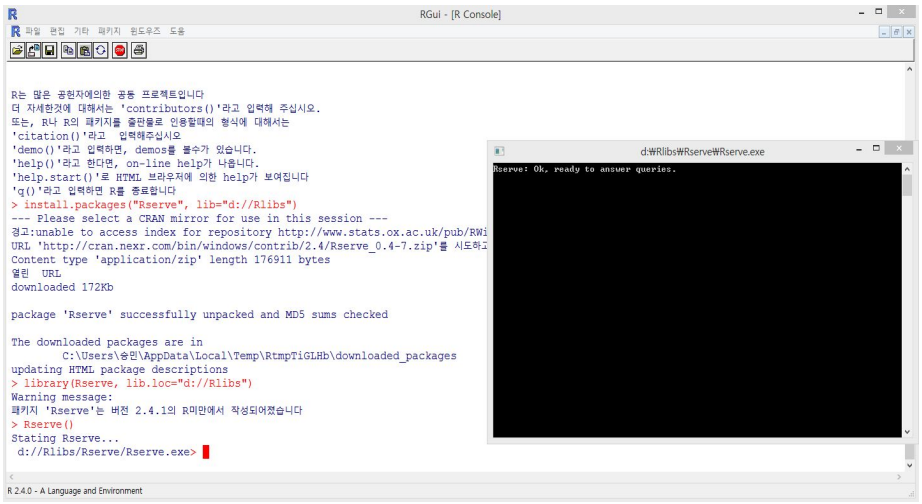
[그림 3-9] 이클립스에서 자바 R 연동 및 실행화면

[그림 3-9]에서 R 3.3.1에서는 라이브러리 JRclient-RF503.jar 이용하여 Rserve() 서버에 접속 할 경우 [Ljava.lang.StackTraceElement;@10b30a7 에러가 발생되었다. JRclient-RF503.jar의 호환 버전이 R 2.4.0 이므로 R 버전을 2.4.0 버전을 다시 재설치하여 진행을 하였으며, 관련 설치법은 R 3.3.1 버전과 동일하게 적용하였다. R 패키지 버전은 R-0.4.9 버전부터 R 3.3.1 까지 지원되는 기능들이 다르기 때문에 사전에 R 홈페이지에서 관련 기능 설명을 반드시 확인하는 것이 필요하다.



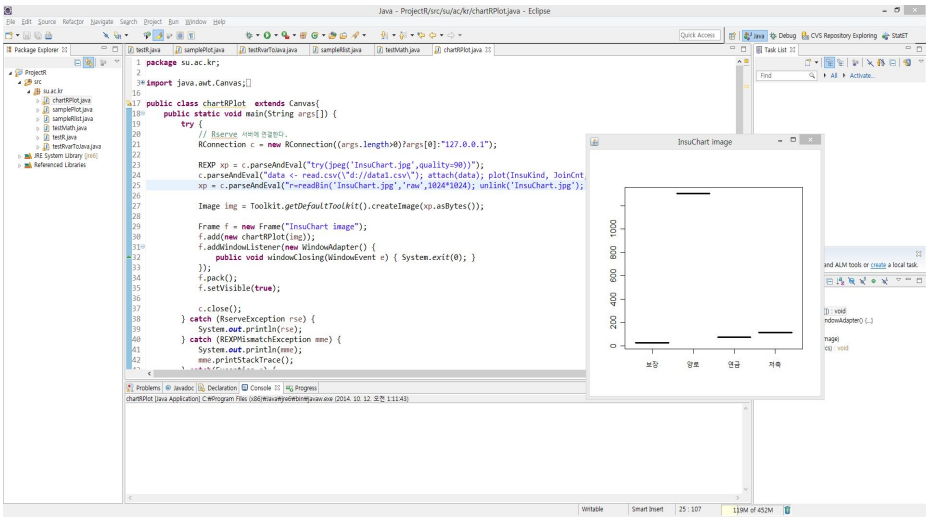
[그림 3-10] R 2.4.0에서 Rserve 서버 설치화면

[그림 3-10] JAVA 연동을 위해 TCP/IP Rserve 서버를 설치하였으며, 기본 포트는 6311 서비스로 가동되고, 윈도우 CMD(Command)를 사용하여 통신 포트 명령어 Netstat로 확인하였다.



[그림 3-11] R 2.4.0 Rserve 서버 실행화면

[그림 3-11]에서는 마찬가지로 R 2.4.0에서 엑셀 패키지 Xlsx를 사용할 때 버전이 충돌하는 현상이 생겨 Read.csv 함수를 사용하였다. R 2.4.0 버전에서 Rserve를 연동할 경우 아래와 같이 정상적으로 CSV 데이터를 가져와서 표 차트가 생성되는 것을 볼 수 있었다.



[그림 3-12] R 2.4.0 JAVA AWT로 구현한 표 차트 생성

[그림 3-12] 화면에서 나타내는 데이터 내용은 다음과 같다.

[표 3-1] 연령별 실적 가입설계 데이터

연령대/보험	보장성	저축	연금	양로
10대	24	111	72	1304
20대	74	786	522	11848
30대	167	1897	1328	20770
40대	203	3430	2362	29436
50대	147	2952	2403	24399
60대	10	158	165	1027

[표 3-1]은 2014년도 연령별 실적 및 보험종류에 대한 데이터를 기록한 것이다.

다음은 보험종류에 대한 간략한 설명이다.

- 보장성보험 : 피보험자에게 사망, 상해, 입원, 생존 등과 같이 사람의 생명과 관련하여 보험사고가 발생했을 때 약속된 급부금을 제공하는 보험 상품으로, 계약 만기 때 지급되는 급부금의 합계액이 이미 납입한 보험료를 초과하지 않는 보험을 말한다.

- 저축보험 : 보장성 보험과 다리 납입보험료보다 만기에 환급되는 보험료가 큰 보험을 말한다.

- 연금보험 : 피보험자의 생존에 대하여 매년 또는 매월 일정액을 지급할 것을 약속한 생존보험을 말한다.

- 양로보험 : 생사혼합보험이라고도 하며, 피보험자가 보험기간 내에 사망하였을 때나, 보험 기간 중 생존해서 만기가 되었을 때에나 다 같이 동일액의 보험금이 지급되는 생명보험을 말한다.

R 데이터를 활용하기 위해서 R의 자료구조에 대한 내용이다.

[그림 3-2] R의 자료구조

Data Structure	Array	Data Type
Vector	1차원	수치/문자/논리/복소수
Matrix	2차원	수치/문자/논리/복소수
Data Frame	2차원	수치/문자/논리/복소수
Array	2차원 이상	수치/문자/논리/복소수
Factor	1차원	수치/문자
Time Series	2차원	수치/문자/논리/복소수
List	2차원 이상	수치/문자/논리/복소수 함수/call/표현식

[표 3-2]에서 R은 사용자의 편의성을 제공하기 위해서 다양한 자료구조를 제공한다. 자바에서 R이 사용하는 자료구조를 사용하려면, Rserve에서 제공하는 REXP 클래스를 사용하면 된다. Rserve에서 사용하는 기본 API(Application Programming InterFace) 중 핵심적인 역할을 수행하는 클래스는 Rconnection 이다. 이 클래스는 R에 접속 후 인증을 수행하고, 세션에 종료, 데이터 파일을 생성/쓰기/읽기 등 각종 자료 전송 및 조회 등을 처리하고 있다.

[표 3-3] Rserve API

생성자	
RConnection()	로컬 호스트에 접속
RConnection(String host)	인자에 입력되는 호스트에 접속
RConnection (String host, int port)	호스트 및 포트를 인자 값으로 사용하여 접속

메소드	
assign	R의변수에 REXP 또는 String 형태로 데이터를 세팅하여 호출
eval	R에 명령을 요청하고 REXP형으로 데이터를 반환
parseAndEval	R서버에 있는 지정 변수의 데이터를 REXP형으로 반환
close	접속을 끊는다.
login	R서버에 해당 계정과 암호로 로그인 한다.

REXP 클래스는 Java와 R에서 서로의 자료구조와 데이터 타입을 전환하여 사용할 수 있도록 지원하는 데이터 모델형 클래스이다. 데이터 프레임과 행렬 구조로 된 데이터모델 등을 생성할 수 있다.

[표 3-4] REXP API

생성자	
REXP()	루트 생성자
REXP(REXPList attr)	REXPList의 REXP 객체를 생성

메소드	
asBytes	Byte 일차원 배열형으로 반환
asDouble	double 형으로 반환
asDoubleMatrix	double 이차원 배열형으로 반환
asDoubles	double 일차원 배열형으로 반환
asList	RList 형으로 반환
asString	String 형으로 반환
asStrings	String 일차원 배열형으로 반환
createDataFrame	RList 형 데이터 리스트를 데이터프레임으로 생성
createDoubleMatrix	double[][] 형 이차원 배열을 R의 행렬형태로 생성
length	데이터의 개수를 확인

Rlist 클래스를 사용하여 Data Frame과 같은 자료 구조를 이용할 수 있는데, Map 인터페이스를 구현하고 있는 Rlist는 내부적으로 Vector 값들을 지닌 리스트를 관리하고 있다.

[표 3-5] RList API

생성자	
RList()	비어있는 리스트를 생성
RList(Collection contents)	Collection형의 리스트를 생성
RList(Collection contents, Collection names)	Collection형의 이름이 있는 리스트를 생성
RList(Collection contents, String[] names)	Collection형, String 배열값의 리스트를 생성
RList(REXP[] contents)	REXP[]형 리스트를 생성
RList(REXP[] contents, String[] names)	REXP[]형, String 배열의 이름이 있는 리스트를 생성

메소드	
at	인덱스 또는 필드명에 해당되는 REXP 객체를 반환
put	Key 데이터 오브젝트를 세팅한다.
putAll	맵(Map)이나 Rlist 형의 모든 데이터를 세팅
remove	Index난 Key에 해당하는 데이터 삭제.
removeAll	Collection에 해당하는 모든 데이터 삭제
size	리스트의 개수를 확인

R의 기본 API 및 JAVA AWT(Abstract Window Toolkit) 활용한 그래프 생성 문법은 다음과 같다.

```
try{
        Rconnection c = new Rconnection();
        System.out.println("===> 접속성공: " + c.isConnected());
}catch(Exception e){
        System.out.println(e.getMessage());
}
```

[그림 3-13] R 접속 테스트

```
REXP xp = c.parseAndEval("try(jpeg('InsuChart.jpg',quality=90))");

c.parseAndEval("data <- read.csv(\"d://data1.csv\"); attach(data);
+ plot(InsuKind, JoinCnt, col=unclass(Species)); dev.off()");

xp = c.parseAndEval("r=readBin('InsuChart.jpg','raw',1024*1024);
+   unlink('InsuChart.jpg'); r");

Image img = Toolkit.getDefaultToolkit().createImage(xp.asBytes());

Frame fr = new Frame("InsuChart image");
fr.add(new chartRPlot(img));
fr.addWindowListener(new WindowAdapter() {
 public void windowClosing(WindowEvent e) { System.exit(0); }
});
fr.pack();
fr.setVisible(true);

c.close();
```

[그림 3-14] R 차트를 생성하는 소스

[그림 3-14]에서 Rserve에서 제공하는 parseAndEval 함수를 이용해 R 서버에서 해당 변수의 데이터를 REXP형으로 반환 받는다. Read.csv 함수를 이용해 CSV데이터를 읽고, 화면크기를 얻을 수 있는 클래스 Toolkit의 getDefaultToolkit을 호출하여 객체를 생성한다. REXP로 반환받는 객체를 이미지객체에 저장을 한다. 마지막으로 이미지를 AWT 프레임 객체에 넣는다. 수행을 마치고 R 서버를 종료한다.

3.2 R과 유사도 알고리즘을 활용한 추천시스템 구현

연령대별 그룹 10대부터 60대까지의 연령별 가입설계 데이터 기반으로 입력받은 사용자A 연령대를 대상으로 가장 유사한 그룹을 찾아서 추천을 하는 시스템을 구현하려고 한다.

사용자A가 입력한 연령대를 토대로 다른 연령대가 가입한 상품 목록

을 A = {a_1, a_2, ..., a_n} 이라고 하자. for a in A 선호도가 높은 a_m을 찾는 방법이다.

코사인 유사도를 사용해서 연령대별 유사도를 계산하였다.

$$\cos(x,y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|}$$

$$x \cdot y = \sum_{k=1}^{n} x_k y_k \|x\| = x \cdot x = \sqrt{(\sum_{k=1}^{n} x_k^2}$$

각 변수의 크기 속성에 ||x||를 계산하고, 속성 값을 ||x||로 나누게 되면, 사용자 A 변수의 보장성보험 속성의 값은 아래와 같이 정규화 할 수 있다.

$$\frac{24}{\sqrt{(24^2 + 111^2 + 72^2 + 1304^2)}} = 0.018$$

따라서 위의 행렬의 속성 값을 정규화한 행렬을 다음과 같다.

[표 3-6] 코사인 유사도를 행렬로 정규화한 데이터

연령대/보험	보장성	저축	연금	양로
10대	0.018307829	0.08467371	0.054923488	0.994725388
20대	0.006225947	0.066129654	0.043918167	0.996824609
30대	0.007990678	0.090768357	0.063542635	0.99381064
40대	0.006828157	0.115372308	0.079448802	0.990116405
50대	0.005723617	0.114939579	0.093563621	0.988939827
60대	0.009504355	0.150168812	0.156821861	0.976097279

이제 위 행렬을 이용하여 사용자간의 코사인 유사도를 계산하였다.
10대와 20대 사이의 코사인유사도는 아래처럼 계산한다.

$$\cos(10대, 20대) = 0.018 * 0.006 + 0.084 + 0.066 ...$$

따라서 사용자간의 유사도 행렬을 아래 표와 같이 계산하여 구할 수 있으며, 대각선을 중심으로는 대칭행렬이므로, 하단의 값은 의미가 없으므로 계산하지 않았다.

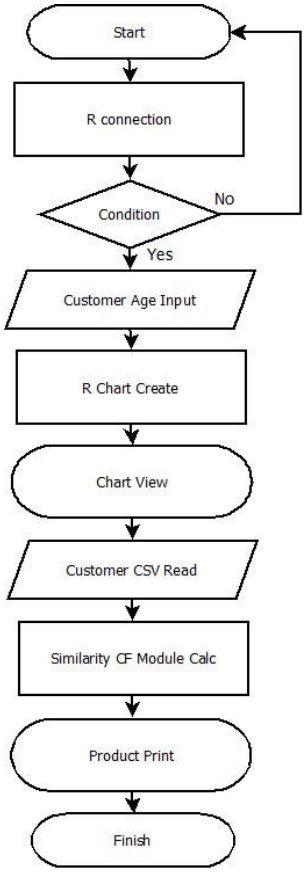
[표 3-7] 연령대 그룹 사이간 코사인 유사도를 대입한 결과

연령대	10대	20대	30대	40대	50대	60대
10대	1	0.9996	0.9998	0.9991	0.9986	0.9923
20대		1	0.9994	0.9981	0.9975	0.9898
30대			1	0.9975	0.9904	0.9936
40대				1	0.9998	0.9962
50대					1	0.9972
60대						1

이제 위의 내용을 토대로 고객의 데이터를 입력 받아 연령대별 그룹과 유사한 연령을 찾는 알고리즘을 R을 활용한 자바로 구현하였다.

[표 3-8] 유사도 알고리즘 구현 소스 및 설명

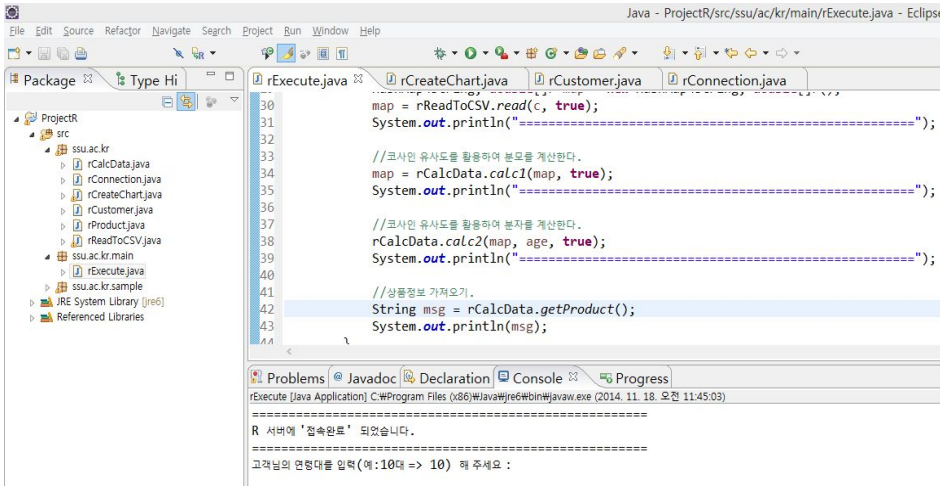
패키지	자바명	내용
ssu.ac.kr	rCalcData.java	코사인 유사도 계산식
ssu.ac.kr	rConnection.java	R 서버 접속
ssu.ac.kr	rCreateChart.java	R 차트 생성
ssu.ac.kr	rCustomer.java	고객정보 입력값
ssu.ac.kr	rProduct.java	상품정보
ssu.ac.kr	rReadToCSV.java	CSV 데이터 읽기
ssu.ac.kr.main	rExecute.java	메인 실행




```
package su.ac.kr.main;
public class rExecute {
public static void main(String[] args) {
//R 서버와 접속을 시도한다.
RConnection c = rConnection.newInstance();

//R 접속 성공시
if(c.isConnected() == true){

//고객 나이를 입력 받는다.
int age = rCustomer.readData();

//차트를 생성한다.
rCreateChart.create(c);

//R패키지에서 CSV 데이터를 읽는다.
HashMap<String, double[]> map =
new HashMap<String, double[]>();

map = rReadToCSV.read(c, false);

//코사인 유사도를 활용하여 분모를 계산한다.
map = rCalcData.calc1(map, false);

//코사인 유사도를 활용하여 분자를 계산한다.
rCalcData.calc2(map, age, false);

//상품정보 가져오기.
String msg = rCalcData.getProduct();
System.out.println(msg);
```

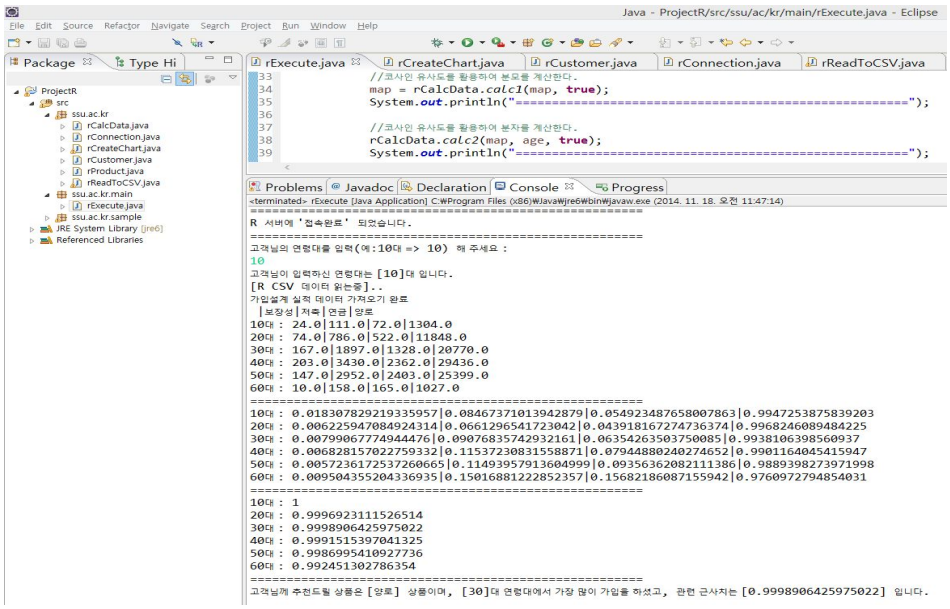
[그림 3-15] 순서도 및 추천 시스템 소스

[그림 3-15]에서 프로세스 순서를 보면, R 서버를 접속을 시도하고, 접속 성공 시, 고객정보 입력 후 R 패키지 API를 사용하여 차트 생성 및 CSV 파일을 읽고, 코사인 유사도 알고리즘을 활용하여 입력 된 고객정보를 비교하여 가장 근사치에 가까운 연령대를 찾아서 실적이 가장 많은 보험 상품을 추천해주는 프로세스를 구현하였다.



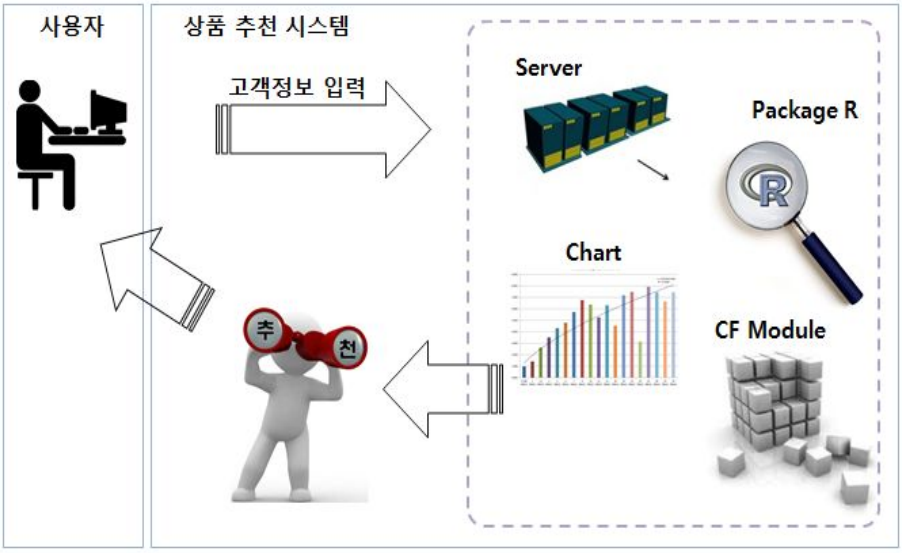
[그림 3-16] 고객정보 입력화면

[그림 3-16] 고객정보 입력을 간략하게 정의하여 10대부터 60대까지 입력 가능하도록 하였다.



[그림 3-17] 결과 값 내용확인

[그림 3-17] 해당 상품에 대한 추천 및 유사한 연령대 그리고 코사인 유사도의 근사치의 값을 확인하였다.



[그림 3-18] R 패키지와 코사인 유사도를 활용한 추천 시스템

3.3 협업적 필터링과 내용기반 필터링 추천시스템의 비교

추천 시스템은 크게 2가지로 나누어서 협업적 필터링과 내용기반 필터링으로 구분 할 수 있다.

(1) 협업적 필터링 (CF : Collaborative Filtering)

협업필터링 상품 추천과정은 입력 데이터, 이웃 집단 탐색, 추천 상품 결정 단계로 나눌 수 있다. 정보의 내용을 직접 분석할 필요 없이 사용자들이 유사도를 측정하여, 상품을 추천할 수 있다.

유사도의 계산은 피어슨 상관계수를 사용하는 Correlation-base 기법, 사용자와 관계있는 이웃을 찾아내는 K-Nearest Neighbor를 이용한

Memory-based CF 기법, Baysian Network, K-Cluster를 사용하는 Model-based CF 기법 등이 있다.

(2) 내용기반 필터링 (CB : Content-based Filtering)

사용자가 선호하는 아이템에 관한 내용을 기록하고, 이와 유사한 아이템을 추천하는 방법으로서 제품 정보를 중심으로 추천하는 방법이다. 적합성 피드백, 가중치 기법, 확률검색 모형, 불리안 검색 등이 활용되고 있으며, 내용기반 필터링은 상품에 포함된 텍스트 정보를 이용하여 필터링 하는데, 이 방법은 논리 연산자와 결합된 문자열 검색 조건에 포함시키거나 삭제해야 될 복잡한 문자열들을 가진 텍스트 프로파일을 사용한다. 텍스트 프로파일에 대한 업데이트 기법은 Genetic Algorithms, Baysian Probality, Machine Learning 기법 등이 활용된다.

[표 3-9] 협업적 필터링과 내용기반 필터링 추천시스템의 비교

구분	협업적 필터링	내용기반 필터링
학습법	이웃 집단(Neighbor Group)	콘텐츠 또는 상품속성
장점	관심사를 실제로 평가 신뢰성 있는 정보 문제를 고려하여 정보공급	상품에 대한 실제평가 추천속도가 빠름
단점	사용자들의 편견 기호의 가치 평준화 효과 데이터의 희박성이 문제 사용자들이 많을 경우 연산속도가 느려지는 현상	유사한 상품을 계속 추천 각 내용을 속성별로 정의하기가 어려움 정확도가 떨어짐
추천방법	사용자가 접하지 못한 상품을 추천	추천 콘텐츠 순위화
보완	차원감소기법	텍스트마이닝 활용

본 논문에서는 두 기법의 혼합형 필터링으로 고객의 연령대 정보를 입력 받은 후에, 코사인 유사도 알고리즘을 활용 후 가중치를 부여, 사용

자의 패턴과 비슷한 이웃 집단(연령대 그룹)을 탐색했다. 또 한 다수 부작용 방지를 위해 모든 유사도의 값의 합으로 나눈 후, 실적이 우수한 상품을 순위화하여 상품을 추천하는 시스템을 구현하였다. 이와 관련하여 좀 더 정확한 상품을 추천하기 위해서는 연령대별 사용자의 선호도에 대한 사항을 프로파일로 구성하고, 상품에 대한 아이템 기반 속성을 좀 더 세분화하여 정확도를 높이는 것이 추가적으로 필요하다. 다음으로는 내용기반 필터링 기법을 사용하여 구현한 유사도 측정 표이다. 각 상품 순으로 상품이 지원하는 속성을 가능 : 1, 불가능 : 2 로 정의하여, 대입하였다.

[표 3-10] 내용기반 필터링 기법을 사용하기 위한 상품속성 표

보험/속성	금리보증	추가납가능	중도인출가능	비과세혜택
보장성	1	1	2	1
저축	2	2	1	2
연금	2	2	2	2
양로	1	1	1	2

[표 3-11] 코사인 유사도를 행렬로 정규화한 데이터

보험/속성	금리보증	추가납가능	중도인출가능	비과세혜택
보장성	0.377964473	0.377964473	0.755928946	0.377964473
저축	0.554700196	0.554700196	0.277350098	0.554700196
연금	0.5	0.5	0.5	0.5
양로	0.377964473	0.377964473	0.377964473	0.755928946

[표 3-12] 상품 사이간 코사인 유사도를 대입한 결과

보험명	보장성	저축	연금	양로
보장성	1	0.838627869	0.755928946	0.571428571
저축		1	0.693375245	0.524142418
연금			1	0.56694671
양로				1

내용기반 아이템 필터링으로 추천할 때는 각 상품에 대한 속성들을 나열하여, 그에 대한 유사도를 가중치를 부여, 사용자가 구매하고 싶은 아이템을 보고 있을 때 근접한 다른 상품을 추천할 수 있다.

[표 3-12]에 도출된 사항으로는 현재 10대 구매 고객이 보장성을 보고 있다고 가정한다면, 그와 비슷하면서 상품 특성이 더 좋은 저축보험을 추천하고, 콘텐츠의 장점을 차트로 비교하여 제공 및 상품을 순위화하여 고객에게 더 나은 선택권을 제공해 줄 수 있다.

제 4 장 결론 및 향후과제

본 논문에서 구현한 추천 시스템은 통계패키지 R과 JAVA AWT GUI를 연동하여 개발자에게는 통계함수에 대한 접근을 용이하게 할 수 있도록 R API를 활용하는 방법을 서술하였으며, 사용자에게는 상품에 대한 시각적인 요소를 제공하기 위해 차트를 출력하고, 문서의 단어의 근사치를 구할 수 있는 필터링 기법 중 하나인 코사인 유사도 알고리즘을 선택하여 구매 이력이 유사한 상품에 대한 근사치를 도출 후 2가지 필터링 기법에 대한 비교를 도출했다.

추천 시스템에서 코사인 유사도 알고리즘을 선택하여 간단한 예측 및 추천은 수행할 수 있었으나 그에 맞는 추천에 따른 기법을 선택해서 사용해야만 사용자에 정보 요구를 효율적으로 해결할 수 있는 것으로 파악된다. 그리고 추천 기법에 대한 부분은 장단점이 존재하기 때문에, 추천의 정확도를 향상 시키려면 필터링의 단점과 데이터 프로파일 구성을 보완 하여 혼합 기법으로 구현하는 것이 좋다.

향후 연구과제는 다양한 기법의 알고리즘의 선정과 시스템 구현 범위에 대한 어려움이 있었는데, 관련연구 조사 및 구현 시스템의 규모가 크게 늘어나는 문제점이 있어 자세하게 다루지 못했다. 여러 가지 설계 모형을 가지고 추천 유형에 맞는 다양한 기법을 연구를 하여, 그에 맞는 필터링 기법을 접목 후 R과 JAVA에 대한 연동 기법 그리고 다양한 인터페이스를 활용한다면 효율적인 추천 기법이 적용된 시스템으로 발전할 수 있을 것으로 기대한다.

참고문헌

[1] 이용필, 숭실대학원 “통계패키지 R과 JSP를 활용한 데이터 분석 시스템 구현“, 2014

[2] 이동우, 제30회 Open Technet 기술세미나, “R, 그리고 빅데이터“, 2014

[3] 주식회사 센소메트릭스, “R의 설치 및 기본 사용법“, 2006

[4] 서민구, (주)도서출판 길벗, “R을 이용한 데이터분석 실무”, 2013

[5] 전희원, NexR Data Scientis, “오픈소스 기반의 통계언어 R과 빅 데이터 분석“, 2012

[6] 유충현, R User Conferenced, ″R의 저변확대를 위한 노력“, 2012

[7] 김용수, 대한산업공학회, “개인화 서비스를 위한 추천 시스템의 연구 동향“, 2012

[8] 김병국, 동국대학교, “유전자 알고리즘을 활용한 개인화된 상품추천 시스템 개발”, 2004

[9] 이준규, 연세대학교, “인터넷 개인화 아이템 추천 알고리즘에 대한 연구”, 2000

[10] 이인기, 용환승, 이화여자대학교, “데이터 마이닝에서 상식을 기반으로 한 유용성 척도”, 2011

[11] IBM, “IBM의 보험 차세대 시스템 솔루션“, 2014

[12] 정태윤, 보험개발원, “통계지표로 보는 우리나라 국민의 생명보험 현황”, 2013

[13] 윤미영, 권정은, 한국정보화진흥원, “빅데이터로 진화하는 세상”, 2012

[14] J H Maindonald, Australian National University, “Using R for

Data Analysis and Graphics Introduction, Code and Commentary", 2008