Өгөгдөл олборлолт, түгээмэл хэрэглээ

Н.Баттүшиг
МУИС, Бизнесийн сургууль
П.Оюунбилэг
МУИС, Бизнесийн сургууль
opagjii44@gmail.com

Н.Мөнхцэцэг
МУИС, Хэрэглээний Шинжлэх Ухаан
Инженерчлэлийн Сургууль
munkhtsetseg@seas.num.edu.mn
Д.Онцгэрэл
МУИС, Бизнесийн сургууль

***Хураангуй*—Энэхүү өгүүлэлд өдөр тутмын үйл ажиллагааны өгөгдөл, мэдээллээс ухаалаг мэдээлэл гарган авах боломжтой өгөгдөл олборлолтын салбарын талаар тодорхой жишээн дээр тайлбарлан авч үзнэ. Бизнесийн байгууллагууд өгөгдөл олборлолтоор үүсгэгдсэн мэдлэгийг цаашдын үйл ажиллагаандаа ашиглаж бизнесээ ухаалагаар удирдаж ашиг олох боломжтой.**

Түлхүүр үг—Өгөгдөл олборлолт, Data mining, алгоритм, машин сургалт, хиймэл оюун ухаан, их өгөгдөл

I. УДИРТГАЛ

Мэдээллийн технологийн эрчимтэй хөгжил, мэдээллийн үнэ цэнэ өсөхийн хирээр хурдацтай өсөн нэмэгдэж буй мэдээллийг хэрхэн боловсруулах асуудал чухлаар тавигдаж байгаа бөгөөд компьютер, сервер, хадгалах төхөөрөмжийн үнэ хямдарч, таблет, ухаалаг гар утаснууд, төрөл бүрийн мэдрэх, хэмжих төхөөрөмжүүд олноор үйлдвэрлэгдэх болсноор цахим өгөгдлийн хэмжээ огцом хурдаар нэмэгдэж байна. Их хэмжээний цахим өгөгдлийг, өгөгдөл олборлолтын аргуудаар боловсруулж бизнесийн үйл ажиллагаанд ашиглах боломжтой [1].

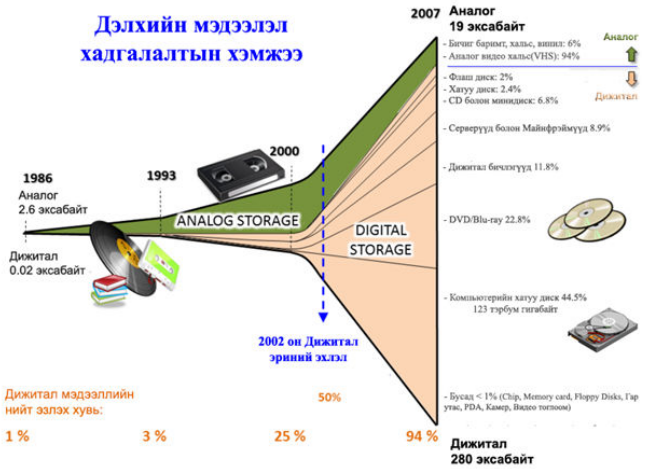
Энэхүү илтгэлийн 1-р хэсэгт илтгэлийн талаарх мэдээлэл, 2-р хэсэгт өгөгдөл олборлолтын тухай, 3-р хэсэгт өгөгдөл олборлолтын алгоритм, процесс, хэрэгсэл, 4-р хэсэгт Microsoft компанийн өгөгдөл олборлолтын хэрэгслүүд, 5-р хэсэгт өгөгдөл олборлолтыг Excel Data Mining Add-in ашиглан бодит жишээн дээр гүйцэтгэж, эцэст нь дүгнэлт, номзүйг оруулсан.

II. ӨГӨГДӨЛ ОЛБОРЛОЛТЫН ТУХАЙ

Өгөгдөл олборлолтын гол зорилго нь: өгөгдлөөс мэдээлэл гарган авах түүнийгээ цаашид ашиглах боломжтой мэдлэг болгон хувиргах юм.

Data - Өгөгдөл (Хадгалагдсан баримтууд, тоо, текст *Жнь, борлуулалтын, мэдрэгчийн, банкны гүйлгээ, макро эдийн засгийн үзүүлэлтүүд, метадата гм*)
Information - Мэдээлэл (Өгөгдлийн зүй тогтол, хоорондын хамаарал зэрэг нь мэдээллийг бүрдүүлнэ. *Жнь: Кассын гүйлгээний өгөгдлийг шинжлэхэд ямар бүтээгдэхүүн, хэзээ зарагдаж байна гэсэн мэдээллийг гарган авч болно*)
Knowledge - Мэдлэг (Мэдээллийг ашиглан түүхчилсэн зүй тогтол, ирээдүйн чиг хандлагын талаарх мэдлэг олж авах боломжтой . *Жнь: Дэлгүүрийн зарагдсан барааны нэгдсэн тайланг ашиглан үйлчлүүлэгчдийн худалдан авалтын зүй тогтлыг илрүүлснээр ямар бараанд урамшуулал үзүүлэхээ тодорхойлох боломжтой.*)

2000 оны үед дундаж компьютер 10GB өгөгдөл хадгалах дисктэй байсан бол өнөөдөр Фэйсбүүкт гэхэд л өдөр бүр 500 терабайт шинэ өгөгдөл нэмэгдэж, Боинг 737 онгоцны нэг удаагийн холын нислэгт 240 терабайт нислэгийн өгөгдөл үүсч, бизнес гүйлгээ, ухаалаг гар утаснууд, ухаалаг мэдрэгчдээр үүсэх өгөгдлүүд нь хэдэн тэрбумаар тоологдох байнга шинэчлэгдэх текст, зураг, дуу, видео гэх мэт төрөл бүрийн хэлбэрийн мэдээллийг үүсгэгдэх болсон [2].



ЗУРАГ №1 ДЭЛХИЙН МЭДЭЭЛЭЛ ХАДГАЛАЛТЫН ХЭМЖЭЭ

Хүн төрөлхтөн өгөгдөл, мэдээллийг боловсруулах тодорхой үе шатуудыг алхан өнөө үед хүрч ирсэн. Хүснэгт №1-д өгөгдөл боловсруулалтын түүхэн үе шатуудыг харуулсан байна.

ХҮСНЭГТ 1 ӨГӨГДӨЛ БОЛОВСРУУЛАЛТЫН ТҮҮХЭН ҮЕ ШАТУУД

Хөгжлийн алхам	Бизнес асуулт	Технологи	Бүтээгдэхүүн үйлдвэрлэгчид	Шинж чанар
Data Collection (1960s) Өгөгдөл цуглуулах	Сүүлийн 5 жилийн миний орлого?	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery Өнгөрсөнд хандсан, тогтмол өгөгдөл
Data Access (1980s) Өгөгдөл боловсруулах	3-р сард Хөвсгөл аймагт бараа хир зарагдсан бэ?	Relational databases (RDBMS), Structured Query Language (SQL) Холбоост өгөгдлийн сан	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level Өнгөрсөнд хандсан, бичлэгийн түвшинд өгөгдлийг уян хатан харуулах
Data Warehousing & Decision Support (1990s) Өгөгдлийн агуулах, Шийдвэр дэмжих	1995 оны 3-р сард Хөвсгөл аймагт бараа хир зарагдсан бэ? Алаг-Эрдэнэ сумыг харуул?	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels Өнгөрсөнд хандсан, олон түвшинд өгөгдлийг уян хатан харуулах
Data Mining (2000 -) **Өгөгдөл олборлолт**	Дараагийн сард Хөвсгөл аймагт хир бараа зарагдах вэ? Яагаад?	Advanced algorithms, multiprocessor computers, massive databases	Oracle, Microsoft, IBM, SAS	Prospective, proactive information delivery Ирээдүйг харсан

Хүснэгт№1-ээс харахад Өгөгдөл олборлолт нь 2000-оноос эрчимтэй хөгжиж эхэлсэн. Data mining (зарим тохиолдолд Knowledge Discovery from Databases – *KDD*) гэдэг нь томоохон өгөгдлийн сангуудаас урьд нь мэдэгдэж байгаагүй тодорхой зүй тогтол бүхий мэдээллийг ялгах (patterns) болон ирээдүйг таамаглах (predict) боломжийг олгодог компьютерийн шинжлэх ухааны нэг салбар юм. Өгөгдөл олборлолт нь нь Хиймэл оюун ухаан (AI) ба Машин сургалт, Статистик, Өгөгдлийн сангийн судалгаа гэсэн салбаруудын огтлолцол дээр оршино. Зураг №2

ЗУРАГ №2 ӨГӨГДӨЛ ОЛБОРЛОЛТ

Ихэнхи бизнесүүд маш их хэмжээний өгөгдлийг олон жилийн ажиллагааны үр дүнд хуримтлуулсан байдаг. Өгөгдөл олборлолтыг ашигласнаар тэрхүү өгөгдлөөс үнэ цэнэтэй мэдлэг гарган авна. Бизнесүүд тэрхүү мэдлэгийг ашиглан, үйлчилгээгээ өргөжүүлэх, борлуулалтаа нэмэгдүүлэх болон ашгаа өсгөх боломжтой. Инженерийн болон анагаахын салбарт энэ нь мөн адил [3].

III. ӨГӨГДӨЛ ОЛБОРЛОЛТЫН АЛГОРИТМ, ПРОЦЕСС, ХЭРЭГСЭЛ

Мэдлэгийг их хэмжээний өгөгдлөөс олборлохын тулд тодорхой аргууд буюу машин сургалтын төрөл бүрийн алгоритмуудыг ашигладаг. Өгөгдөл олборлолтын алгоритм гэдэг нь өгөгдлөөс олборлолтын загвар үүсгэх бүлэг тооцооллуудыг нэрлэдэг. Түгээмэл хэрэглэгддэг алгоритмуудыг төрлөөр нь нэгтгэвэл:

Classification (Ангилалт) – Өгөгдлийг ялгаатай ангиудад хуваахад ашиглагддаг. Түгээмэл хэрэглэгддэг алгоритмууд нь naïve Bayes, decision tree, regression, neural network зэрэг алгоритмууд ордог. (supervised)

Жнь: Үйлчлүүлэгчдийн дундаас өрсөлдөгч компани руу шилжих магадлалтай хэсгийг ангилах(churn modeling)

Clustering – (Бүлэглэлт) Төсөөтэй шинж чанар бүхий обьектуудыг бүлэглэх алгоритмууд.(unsupervised)

Жнь: Ижил худалдан авалтын хэвшилтэй хэрэглэгчдийг илрүүлэх

Prediction (Таамаглалт) – Мэдэгдэхгүй байгаа утгыг өгөгдлийн бусад талбаруудаас хамааруулан таамаглана. Хэрвээ таамаглах хувьсагчийн утга, тасралтгүй байвал регрессийн алгоритмууд өргөн хэрэглэгддэг.

Жнь: Борлуулалт ирэх саруудад ямар байхыг таамаглах

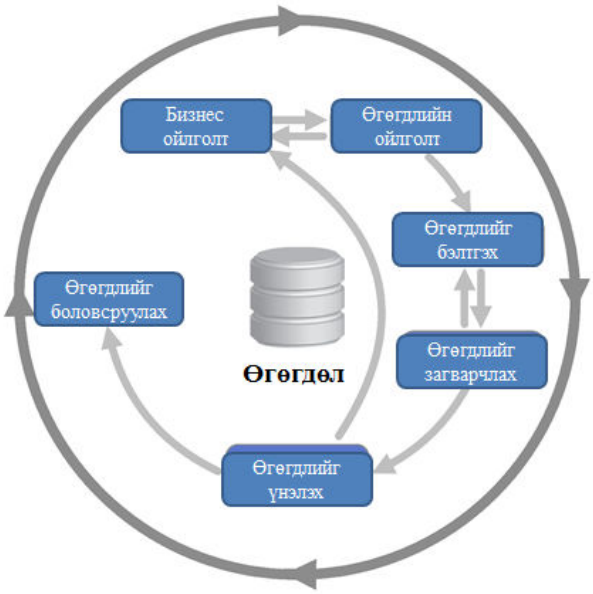
Association (Хамаарал) – Өгөгдлийн дунд байх ялгаатай шинж чанаруудын хоорондын хамаарлыг хайж олно. Хамгийн түгээмэл хэрэглээ нь худалдан авалтын сагсны шинжилгээнд ашиглагдах хамаарлын зүй тогтлын илрүүлэлт юм.

Жнь: Худалдаж авах(х, талх)->Худалдаж авах(х, сүү)

Бүлэглэлтийн алгоритм нь “unsupervised” хэлбэрт хамаардаг бөгөөд энэ хэлбэрийн алгоритмууд нь нийт өгөгдөлд байгаа бүхий л зүй тогтлуудыг шинжилдэг. Харин ангилалтын алгоритмыг агуулах “supervised” хэлбэрийн алгоритмууд нь зөвхөн урьдчилан тодорхойлсон хамааран хувьсагчдад нөлөөлөх зүй тогтлуудыг шинжилдэг [4].

Олборлолтын загвар үүсгэхийн тулд алгоритм эхлээд туршилтын өгөгдөлд шинжилгээ хийж тодорхой хэв маяг, хандлагыг тодорхойлно. Алгоритм шинжилгээнийхээ үр дүнг олон давтах замаар олборлолтын загварын оновчтой параметрүүдийг тооцоолно. Үүний дараагаар уг загварыг цуглуулсан нийт өгөгдөлд биелүүлснээр үр ашигтай байж болох зүй тогтол, нарийвчилсан статистикуудыг илрүүлнэ.

Өгөгдөл олборлолт гүйцэтгэхэд хамгийн түгээмэл хэрэглэгддэг процессийн загвар нь Cross Industry Standard Process for Data Mining (CRISP-DM) “Зураг №3”юм.



ЗУРАГ№3 CRISP-DM процессийн диаграм

CRISP-DM процесс нь дараах алхмуудаас тогтоно.

- Бизнес ойлголт (Business Understanding) – Төслийн зорилгыг бизнесийн талаас нь тодорхойлж, түүнийгээ өгөгдөл олборлолтын асуудал болгон тавих
- Өгөгдлийг ойлгох (Data Understanding) – Өгөгдөлтэй танилцаж, чанарыг нь шалгах, таамаглал дэвшүүлэх боломж бүхий дэд хэсгүүдийг ялгах
- Өгөгдлийг бэлтгэх (Data Preparation) – Дараагийн шат болох өгөгдлийг загварчлахад шаардлагатай эцсийн өгөдлийн санг гарган авах
- Загвар байгуулах (Modeling) – Төрөл бүрийн загварчлалын арга техникүүдээс сонгон хэрэгжүүлж, тэдгээрийн параметрүүдийг оновчтой үр дүн гаргахаар тохируулна. Ер нь нэг төрлийн асуудалд хэд хэдэн ялгаатай техник хэрэглэгддэг.
- Үнэлэлт (Evaluation) – Үүсгэсэн загварыг дахин шалгаж, турших, бизнес зорилгуудыг хэрхэн хангаж байгааг нягтлах
- Эцсийн хэрэглэгчдэд хүргэх (Deployment) – Олборлолтын загвар үүсгэснээр төсөл дуусдаггүй, түүнийг хэрэглэснээр гарсан үр дүнг хэрэглэгчдэд ойлгомжтой байдлаар хүргэх шаардлагатай[5].

Орчин үед өгөгдөл олборлолтыг гүйцэтгэх төлбөртэй болон төлбөргүй олон програм хангамжууд бүтээгдсэн ба өгөгдөл олборлолтод ашиглаж болох нээлттэй эхийн програм хангамжуудыг жагсаавал:

- R-Programming – Статистикийн тооцоолол болон графикт зориулсан програмчлалын хэл, програмчлалын орчин
- Orange – Програмчлалын Python хэл дээр суурилсан нээлттэй эх бүхий хэрэгсэл, биоинформатик болон текст анализ гэх мэт олон нэмэлт боломжуудтай
- WEKA – Java хэл дээр суурилсан олон төрлийн алгоритм, шинжилгээний аргуудыг агуулсан хэрэгсэл
- RapidMiner Basic – Өгөгдөл олборлолтын иж бүрдэл програм, мэргэжлийн хувилбарууд нь төлбөртэй.
- KNIME – Өгөгдлийн шинжилгээ хийх хүчирхэг хэрэгсэл, Eclipse, Java

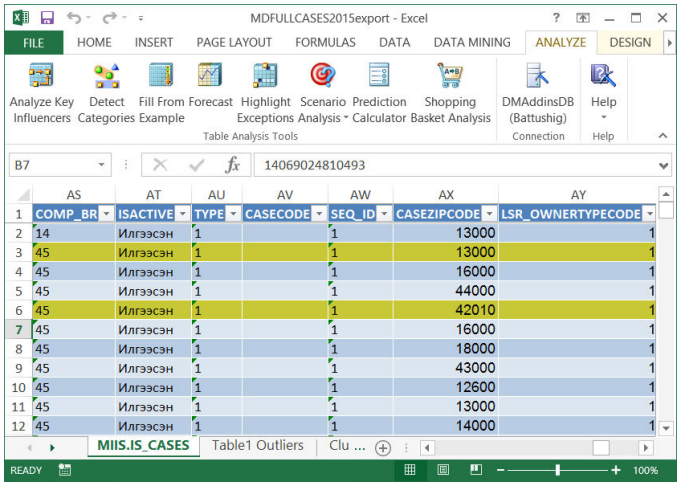
Төлбөртэй өгөгдөл олборлолтын хэрэгслүүд нь өндөр үнэтэй байдаг нь бодит өгөгдөлд олборлолт хийж, бизнесийн үйл ажиллагаанд ашиглахад хүндрэлтэй болгодог.

IV. MICROSOFT КОМПАНИЙН ӨГӨГДӨЛ ОЛБОРЛОЛТЫН ХЭРЭГСЛҮҮД

Microsoft компани нь SQL Server өгөгдлийн сан удирдах системийн бүрэлдэхүүндээ өгөгдөл олборлолтын алгоритмуудыг оруулан хөгжүүлж ирсэн ба SQL Server Analysis Service нь өгөгдлийн олборлолт, шинжилгээг гүйцэтгэдэг. SQL Server-г шууд ашиглан шинжилгээ олборлолт хийх нь хүндрэлтэй, програмчлалын мэдлэг, туршлага ихээхэн шаардагддаг.

Иймд Microsoft компани SQL Server 2008 хувилбарыг дэмжин ажиллах өгөгдлийн олборлолтын нэмэлт хэрэгслийг Excel 2007 хувилбарт зориулан бүтээж, дараа дараагийн хувилбаруудад улам сайжруулсаар ирсэн. Энэхүү нэмэлт хэрэгсэл нь төлбөргүйгээс гадна Excel програмын хүснэгт, графикийн мэдээлэлтэй ажиллах, тооцоолох чадвартай хослсноор зах зээлд борлуулагдаж буй өгөгдлийн олборлолтын програмуудтай өрсөлдөхүйц бүтээгдэхүүн болж чадсан.

Excel дэх өгөгдөл олборлолтын хэрэгслүүд нь Analyze, Data Mining гэсэн хоёр цэсэнд хуваагдан байрладаг. Analyze буюу хүснэгт анализийн хэрэгслүүд нь тодорхой үйлдлүүдийг гүйцэтгэхэд зориулагдсан 8 командыг агуулдаг. Эдгээр хэрэгслүүдийг ашиглахад түүнд хэрэглэгдэж байгаа өгөгдөл олборлолтын алгоритмуудыг заавал мэдэх шаардлагагүй юм.



ЗУРАГ№4 ANALYZE ЦЭСНИЙ КОМАНДУУД

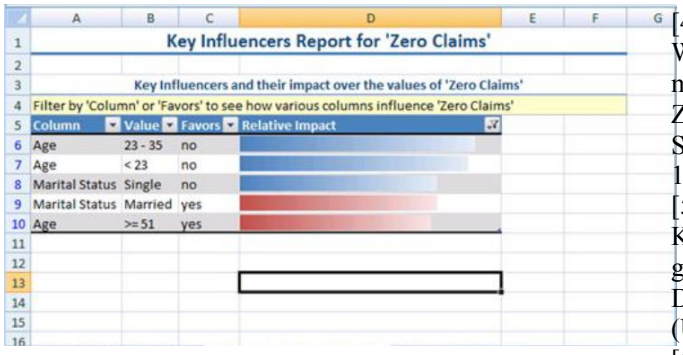
Data Mining бүлэг нь өгөгдөл олборлолтын талаарх мэдлэгтэй хэрэглэгчдэд зориулагдсан ба SQL Server Analysis Service-н бүхий л боломжуудыг ашиглах боломжийг олгоно. Мөн хэрэглэгчид алгоритмын параметрүүдийг өөрчлөх боломжтой тул хүснэгт анализийн хэрэгслээс илүү уян хатан, үр дүнтэй байдаг.

V. EXCEL DATA MINING ADD-IN АШИГЛАХ

Даатгалын салбар нь өгөгдлийн олборлолт өргөнөөр ашиглагддаг салбаруудын нэг юм. Автомашины даатгалын хувьд ямар үйлчлүүлэгчдийн хувьд эрсдэл багатай, нөхөн төлбөр авдаггүй болохыг тогтоож түүнд даатгуулагчийн нас, туршлага, тээврийн хэрэгслийн загвар, даатгалын хураамж гэх мэт шинж чанарууд ямар нөлөөтэй байдгийг илрүүлэх нь чухал ач холбогдолтой.

Excel өгөгдөл олборлолтын хэрэгслийн Analyse Key Influencers команд нь хүснэгтийн нэг талбарын

утганд бусад талбарууд хэрхэн нөлөөлдөгийг харуулдаг тул түүнийг ашиглан эрсдэлгүй даатгуулагчдын шинж чанаруудыг таамаглав.



	A	B	C	D	E	F	G
1	Key Influencers Report for 'Zero Claims'						
2							
3	Key Influencers and their impact over the values of 'Zero Claims'						
4	Filter by 'Column' or 'Favors' to see how various columns influence 'Zero Claims'						
5	Column	Value	Favors	Relative Impact			
6	Age	23 - 35	no				
7	Age	< 23	no				
8	Marital Status	Single	no				
9	Marital Status	Married	yes				
10	Age	>= 51	yes				
11							
12							
13							
14							
15							
16							

ЗУРАГ№5 ANALYZE KEY INFLUENCERS ХЭРЭГСЛИЙН УР ДҮН

Алгоритмын үр дүн нь өгөгдөлөөс шууд хамаардаг бөгөөд Analyze Key Influencers хэрэгслийн үр дүнгээс харахад гэрлэсэн, 51-с дээш насныхан эрсдэлгүй буюу нөхөн төлбөр аваагүй байна. Ганц бие, ялангуяа 35-с доош насныхан эрсдэлтэй буюу даатгалын тохиолдол гаргаж, нөхөн төлбөр авсан байх магадлал өндөртэй нь харагдаж байна.

ДҮГНЭЛТ

Цахим өгөгдлийн хэмжээ жил бүр огцом нэмэгдэж байгаа нь уг өгөгдлийг олборлох замаар оновчтой шийдвэр гаргах, бизнесийн үйл ажиллагааг сайжруулах боломжийг олгодог. Өгөгдөл олборлолт нь машин сургалтын алгоритмуудыг бодит өгөгдөлд нэвтрүүлж үр дүн гарган авах үйл ажиллагаа юм. Өгөгдөл олборлолтын алгоритмуудын чанар сайжирсаар байгаа боловч тодорхой шинжилгээний асуудалд хамгийн сайн тохирох өгөгдлийн олборлолтын алгоритмыг сонгох нь хүндрэлтэй байдаг.

Microsoft компанийн Excel Data Mining Add-In-г ашигласанаар хэрэглэгчдийн хэзээний танил интерфэйсийг ашиглан SQL Server-т агуулагдах өгөгдөл олборлолтын хүчирхэг алгоритмуудыг хялбараар ашиглах боломжтой.

АШИГЛАСАН МАТЕРИАЛ

[1] Mining Business Databases. 42-48, s.l. : Communications of the ACM, Vol. 39(11).

[2] Paul Zikopoulos, Chris Eaton. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. s.l. : McGraw-Hill Osborne Media, 2011 . ISBN:0071790535 9780071790536 .

[3] Link Mining: A New Data Mining Challenge. Getoor, L. 84-89, s.l. : SIGKDD Explorations, 2003, Vol. 5(1).

[4] Top 10 algorithms in data miningKnowl Inf Syst. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg. Knowl Inf Syst : Springer-Verlag London Limited , 2008, Vols. 14:1–37.

[5] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR). CRISP-DM 1.0 Step-by-step data mining guide. CR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) : SPSS inc, 2000.

[6] Ville, Barry de. Microsoft Data Mining Integrated Business Intelligence for e-Commerce and Knowledge Management. s.l. : Elsevier Inc, 2001. ISBN: 978-1-55558-242-5.