Мэдээллийн интеграцийг онтологи ашиглан зохион байгуулах нь

Ш.Бат-Өлзий
Монгол Улсын Их Сургууль
Мэдээллийн системийн тэнхим
Д.Гармаа
Монгол Улсын Их Сургууль
Програм хангамжийн тэнхим

ABSTRACT

Ontology and data integration is one of the main direction of modern artificial intelligence systems. In this frame we are seeking new methods and tools for web data integration using OWL and ontology languages. Our main task is to create and use midlleware system for data integretion.

KEYWORDS

Data integration, OWL, Ontology, Semantic web

1. ҮНДСЭН АСУУДАЛ

Мэдээллийг хэрэглэж буй байгууллагууд тухайн мэдээллээ чөлөөтэй мөн тархмал байдлаар хэрэглэх шаардлагыг тавьж байдаг. Ихэнхи хэрэглэгчид буюу мэдээллийн системүүд нь дээрхи зорилгынхоо үүднээс өөрийн гэсэн загвар форматтай өгөгдөл, өгөгдлийн санг хэрэглэж байдаг. Энэ нь нэг систем нөгөө хэрэглэгчийн буюу системийн хэрэглэж байгаа мэдээллийг хэрэглэх шаардлагатай болбол уг өгөгдлийг өөрийн системийн тохирох форматад хөрвүүлэх үйл ажиллагааг хийх шаардлагатай болно. Энэ үйл ажиллагаа нь ихээхэн цаг шаардсан өртөг зардал өндөртэй ажиллагаа болдог.

Өгөгдлийг олж эзэмших үйл ажиллагаа нь хагас автомат, бүрэн автомат эсвэл хэрэглэгчийн үйл ажиллагаагаар хангагдаж байдаг. Өгөгдлийг хамтран эзэмших үйл ажиллагаа гэдэг нь оршин байгаа өгөгдлийн эх үүсвэрээс дээрхи асуудлыг шийдвэрлэхээр мэдээллийг хувиргаж байх үйл ажиллагаа гэж үзэж болно.

Дээрхи асуудлыг шийдвэрлэхийн тулд олон тооны технологи, аргууд хөгжүүлэгдсэн байна. Эхэндээ мэдээллийн тохирох эх үүсвэр нь тухайн хэрэгцээг хангаж байх мэдээллийг яг агуулсан байх шаардлагатай гэж үздэг байсан байна.
Мэдээллийг хайх явцад нэг л удаа мэдээллийн эх үүсвэр олдсон тохиолдолд түүнд байгаа мэдээллийг ашиглах шаардлагатай болно. Тархалттай компьютерийн системийн хувьд өгөгдлийн утга зүйн тэнцүү чанар ихээхэн асуудал үүсгэдэг. [1]

Энэ төрлийн системүүдийн хувьд судлаачид онтологийг түлхүүр технологиор сонгон авсан байдаг. [2] Хамгийн сайн загварчлагдсан системүүдийн хувьд өгөгдөл нь өгөгдлийн сангаас түүнд ашиглагдаж байгаа онтологи руу дамжуулагддаг байна. Энэхүү онтологид суурилсан системийг нь Intelligent Information Integration [3] хэмээн нэрэлсэн байна.

2. ТЕХНОЛОГИ

Мэдээллийн өргөн орчинд нэгэн төрлийн бус өгөгдөл мэдээллийг хэрхэн нэгэн утгатай ашиглах вэ гэдэг нилээн ярвигтай асуудал байдаг. Ерөнхийдөө нэгэн төрлийн бус орчны асуудал нь үндсэн 3-н төрөлд хуваагдна гэж үздэг.

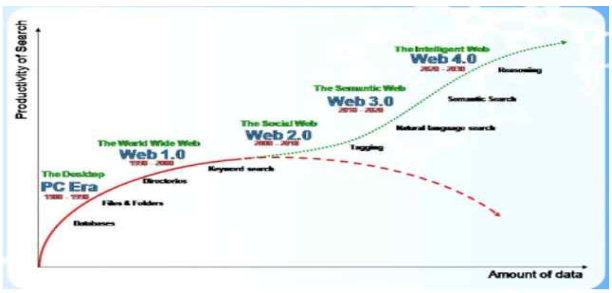
1. Синтакс (өгөгдлийн загварын нэгэн төрлийн бус байдал)
2. Бүтэц
3. Семантик гэсэн үндсэн 3 төрөлд хуваагдна гэж үзсэн байна.

Энэхүү хэсэгт бид семантик нэгтгэл болон агуулгад суурилсан шүүлт хийх, тэдгээрийг хэрэгжүүлэх аргуудыг хөгжүүлэх тухай судлан үзэх болно. Синтаксийн түвшинд стандартчилал нь хамгийн гол чухал хэсэг хэмээн үздэг. Ялгаатай өгөгдлийн эх үүсвэрүүдийг нэгтгэхэд хэрэглэгдэхээр олон тооны стандартууд хөгжүүлэгдсэн байна. Хамгийн сонгодог өгөгдлийн сангийн интерфэйс ODBC ээс гадна веб хандалтад HTML [Ragget ба бусад., 1999], XML [Bray ба бусад., 1998] ба RDF [Lassila ба Swick, 1999] гэсэн технологиуд гарч ирсэн байна.

Компьютерийн сүлжээ өргөжин тэлэхийн хирээр W3C-ийн гаргаж буй стандарт хэлүүд ихээхэн хурдтайгаар шинэчлэгдэн гарч байна. Эдгээрийн зарим стандартууд нь мэдээллийн нэгтгэл, хамтран эзэмшилд зориулагдаж байна. Манай гол анхаарал нь мэдээллийн өргөтгөсөн тэмдэглэлийн хэл XML ба нөөцийг тодорхойлох загварын хэл RDF байх болно.

3. ШИЙДЭЛ

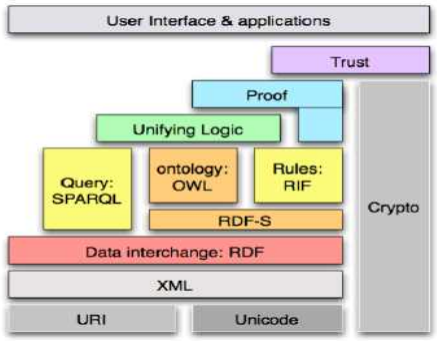
Вебийн технологийн хөгжлийн үйл явц болон үе шат бүрт хэрэглэгдэх технологиудын талаар дараах 2 зургаар харуулав.



Зураг №1 Вебийн хөгжлийн үе шат болон хөгжлийн хугацаа

Дээрхи зургаас харахад 2010 оноос эхлэн семантик веб буюу Web 3.0-ийн хөгжлийн болон хэрэгжилтийн хугацаа эхэлсэн байна. Энэхүү үе шатуудад мэдээллийн хэмжээ болон технологийн хоорондын хамаарал урвуу болж байгааг харж болно.

Өөрөөр хэлбэл тухайн системүүд буюу веб системүүд хэрэглэж байгаа технологио хэрхэн өөрчлөхөөс ихээхэн хамаарна. Технологио өөрчлөхгүй бол өгөгдлийн хэмжээг удирдах боломжгүй болно.



Зураг № 2 Семантик вебийн архитектур

Вебийн технологийн хувьд семантик вебийн технологийн стандартыг хамгийн сүүлийн байдлаар W3C –ээс баталсан байна. Уг стандартыг дээрхи зургаар харуулав.

Хамгийн анхын асуудал бол түүнийг тодорхойлох үе шатанд гарах ба нийлмэл бүтэцээс хамаардаг байна. Урт хугацааны турш энэ асуудал хөрвүүлэлтийн хүсэлтийг гараар бичих замаар шийдвэрлэгддэг байсан байна. Энэ нь ялгаатай өгөгдлийн загваруудын хооронд солилцоог хийж байдаг дундын орчны системийн талаархи судалгаа, шинжилгээг ихээхэн нэмэгдүүлсэн байна.

Энэхүү ялгаатай өгөгдлийн загваруудын хооронд мэдээлэл солилцох дундын систем нь ялгаатай өгөгдлийн бүтцүүдийн хооронд мэдээлэл солилцох үндсэн дүрмийг бий болгох шаардлагатай юм. Ялгаатай бүтэцтэй өгөгдлийн сангуудын хооронд мэдээлэл солилцох энэхүү асуудлыг шийдвэрлэхээр олон судлаачид тус тусын хувилбаруудыг дэвшүүлсэн байна

Схем(Schema), Хүсэлт(Query) ба Мэдээллийн шүүлт(View)

Схемийг нэрлэгдсэн холбоосуудын олонлог гэж энгийн байдлаар тодорхойлж болно. Холбоос дахь байрлалуудыг багана гэж нэрэлдэг ба тодорхой нэрээр нэрлэж

схемийн үндсэн хэсэг гэж тооцдог. Холбоост өгөгдлийн мэдээллүүдийг ашиглах хэлбэр нь дараах байдлаар бичигднэ.

$$q(\overline{X}) \Leftarrow e_1(\overline{X}_1) \wedge \ldots \wedge e_n(\overline{X}_n)$$

$e_1 \ldots e_n$ нь холбоосууд, $\overline{X}_n$ нь тухайн холбоос дахь элементийн хослолууд нв байна. Энэхүү query ийн үр дүн нь хос бүрийн олонлог байна. Мэдээллийн шүүлт гэдэг нь нэрлэгдсэн хүсэлтийг хэлнэ. Олон төрөлт байдалтай өгөгдлийн схемүүдийг хооронд нь нэгтгэхэд олон төрөлт схемүүдтэйгээ дугаарлагдсан шүүлтээр холбогдсон глобал схемийг хэрэглэнэ. Дараах 2 төрлийн үндсэн хандлага байдаг гэж үзсэн байна. Энэ хандлага нь схем хоорондын хөрвүүлэлтийн чиглэлээс хамаардаг байна.

- ***Глобал шүүлтийн хандлага***

Глобал шүүлтийн хандлага нь глобал схем дахь холбоос бүр нь ялгаатай схемүүдэд нэгтгэгдэх боломжтой шүүлтүүдээр тодорхойлогдно. Ялангуяа нэгтгэл нь салаалсан дүрмээр тодорхойлогдно. Энэ дүрэм нь ялгаатай өгөгдлийн олон төрөлт байдалтай схем дэх холбоосуудын эхний таарах хэсэг ба нэгтгэгдсэн схем дэх холбоосын үр дүнгээр тодорхойлогдно. Глобал схемд тавигдсан хүсэлт(query) нь шүүлтийн тодорхойлолтуудыг ашиглан нөхцлүүдийг нээх замаар хариултаа авна. Нээгдэж байгаа хүсэлт нь өгөгдлийн сангийн системд хэрэглэгдэж байгаа энгийн уламжлалт техникийг ашиглан тодорхойлогдно. Энэ хандлагын дутагдалтай тал нь тус тусдаа саланги мэдээллийн системүүдийн бие даасан байдал нь нэгтгэгдсэн нэг хүсэлтийг бий болгоход алдагдах явдал юм. Үүнийг шийдвэрлэх асуудал нэгтгэгдсэн системд мэдээллийн эх үүсвэрийг нэмэх болон хасах үед шаардлагатай болно.

- ***Локал шүүлтийн хандлага***

Локал хүсэлтийн хандлага нь шүүлтүүд нь эсрэг замууд дээр хэрэглэгдэх байдлаар тодорхойлогдно. Өгөгдлийн хувиргалтыг гүйцэтгэх нь мэдээллийн нэгтгэлийн хувьд чухал үүрэгтэй байна. Мета өгөгдлийн хувьд өгөгдлийн хувиргалт нь өгөгдлийн эх үүсвэрийн загвараас хүсэлт гаргасан хэрэглэгчийн өгөгдлийн загварт шилжүүлэх үйлдэл юм. өгөгдлийн хувиргалт нь үндсэн 2 алхамд хуваагдна.

1. Өгөгдлийн элементийг өгөгдлийн эх үүсвэрээс хүлээн авагч руу өгөгдлийг буулгах ба ямар нэгэн хувиргалтыг заавал гүйцэтгэнэ.
2. Кодын үүсгүүр нь хувиргалтын бэлэн програмыг үүсгэнэ.

Тодорхойлолт 1 Хэлний хамрах хүрээ

L' гэсэн хэлийг дараах нөхцлийг хангасан $\tau : L \to L'$ буулгалт оршин байвал L хэлийг агуулж байна гэж хэлнэ.

(4.1) $\exists \tau, (\forall \delta \in L, \tau(\delta) \in L')$

Бид энэ агуулагдлыг $L < L'$ гэж тэмдэглэнэ.

Бид L'-ийн орчинд мэдлэгийг хамтран эзэмших хэд хэдэн аргыг тодорхойлсон бөгөөд эдгээр нь дараах үндсэн аргууд байна гэж үзлээ.

Буулгалтын арга- хамгийн өргөн хэрэглэгддэг энэ арга нь өгөгдлийн үүсвэр хэлний илэрхийллийн төрлүүдийг хувиргах хэлнийхээ элементүүдэд хувиргаж шийдвэрлэнэ. Энэ арга нь нэг хэл нөгөө хэлэндээ хувирсан бол уг холбоосоор буцах холбоос болон хувирна. Энэ нь ихээхэн ашиггүй арга бөгөөд хөрвүүлэлтийн атрибутыг тусгайлан шалгадаг тул хурдны хувьд удаан юм.

Үндсэн хэсгийн арга Энэ арга нь хөрвүүлэлтийн тоог багасгахын тулд тухайн тухайн хэлнүүдэд тохирсон тусгай хөрвүүлэлтийн бүтцийг ашигладаг арга юм. Хамгийн ерөнхий бүтэцтэй энгийн Lp гэсэн хэлийг үүсгэх ба бусад хэлнүүд нь энэ хэл рүүгээ хөрвүүлэгднэ. Энэ үндсэн хэл нь дараах байдлаар хязгаарлагдна.

$$\exists! L_p, \forall L_i (L_i \prec L_p)$$

Түвшний арга энэ арга нь хэлнүүдийг агуулж байгаа түвшнүүдийн хооронд семантик харилцан үйлдлийг гүйцэтгэдэг. Энэ аргын нэг жишээ нь OIL хэл юм. [Fensel ба бусад, 2000]. Энэ хэл нь өнөөдөр WWW-ийн стандартад орсон байна. Энэ аргын гол санааг

$$\forall i, j, (i \le j \Rightarrow L_i \prec L_j)$$

гэсэн томъёогоор тодорхойлж болно.

4. МЭДЭЭЛЛИЙН ИНТЕГРАЦИ ТҮҮНИЙ ЭЛЕМЕНТҮҮД

Өмнөх хэсгүүдэд бид мэдээллийн интеграцийн framework систем, түүний онолын үндсэн ойлголтуудын талаар судлаж үзсэн билээ. Өнөөдөр ихэнхи судлаачид өгөгдлийн интеграцийн системийн формалчилалыг тодорхойлсон Lenzerini-ийн загварыг мөрдөж байна. Энэ загвар нь мэдээллийн интеграцийн систем нь (G,S,M) гэсэн гурвалаас тогтно гэж үздэг.

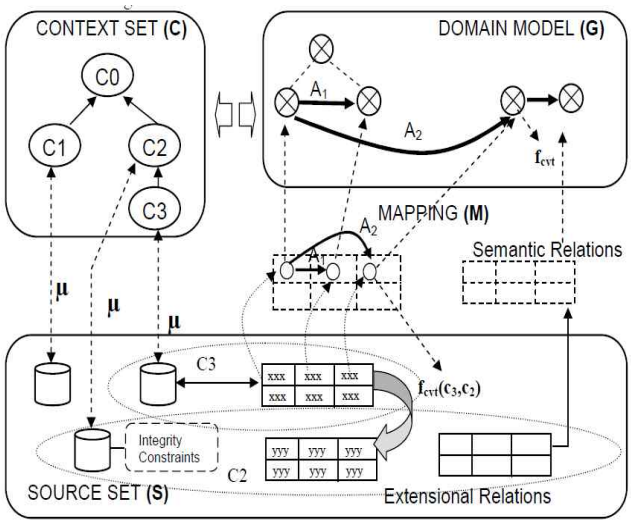
- G- глобал схем, домайн модел, онтологийг илэрхийлнэ.
- S – мэдээллийн эх үүсвэр
- M - G ба S-ийн хоорондох хувиргалт [4]

Бид энэ системийг судласны үндсэн дээр энэ системийг дараах байдлаар өргөтгөх боломжтой хэмээн хувилбар дэвшүүлж байна. Мэдээллийн интеграцийн өргөтгөсөн систем нь

(G, S, M, C, μ) гэсэн хэлбэрээр тодорхойлогдно. Үүнд:

- C - агуулгын нийлмэл олонлог
- μ - агуулгыг өгөгдлийн эх үүсвэртэй холбох буулгалт

Зураг 3-т бид энэхүү өргөтгөсөн системийн элэментүүдийн хоорондын харилцан хамаарлыг харуулав. Дугуй өнцөгтэй тэгш өнцөгтөөр G, S, C- гэсэн элементүүдийг дүрслэн харуулав.



Зураг№3 Өгөгдлийн интеграцийн элементүүдийн хоорондын хамаарал

6. ДҮГНЭЛТ

Судалгааны ажлын хүрээнд нэгэн төрлийн бус өгөгдлийн орчинд мэдээллийн интеграцийг хэрхэн гүйцэтгэх тухай асуудлыг авч үзэж байна. Мэдээллийн системийн интеграцийг гүйцэтгэхдээ DAML+OIL, OWL хэлийг ашиглан ontology –д суурилсан өгөгдлийн эх үүсвэр болон хувиргах эх үүсвэрийн хооронд мэдээллийг хувиргах, мета өгөгдлийг тодорхойлох үйлдлийг зохион байгуулах шаардлагатай юм. Мэдээллийн интеграци системийг зохион байгуулахад вебийн орчний болон application орчны гэсэн 2 хувилбараар хийгддэг ба Oracle ETL болон MSSQL ийн интеграци зэргээр хийгдэх боломжтой.

6. НОМ ЗҮЙ

[1] Д. Гармаа, Д. К. Вэйк – “ Мэдээллийн нэгдсэн сүлжээнд холбоост өгөгдлийн сангуудыг нэгтгэх ”, 2003

[2] Fensel, 2001, Gruninger and uschold, 2009 “Data integration and semantics”

[3] R.Karp, V.Chaudhri, and J.Thomere “XOL and XML-based ontology exchange language” , Aug, 2010, www.ai.sri.com

[4] S.Lukeand J.Heflin “*SHOE 1.01 proposed specification* SHOE project, Feb, 2000

[5] R.Kent, Conceptual Knowledge Markup Language,2008, www.ontologos.org/ckml/ckml2002.html