Шийдвэрийн Модны Ангилагч Ашиглан Монгол Зохиолын Зохиолчийг Тодорхойлох нь

Дамиран Золбоо, Алтангэрэл Хүдэр
Шинжлэх Ухаан Технологийн Их Сургууль, Компьютерийн Техник Менежментийн Сургууль, Компьютерийн Ухааны Салбар
d.zolboo@csms.edu.mn, khuder@csms.edu.mn

Хураангуй **– Зохиолчийг тодорхойлох асуудлын үндсэн зорилго нь зохиогчийн эрхийг зөрчих, хэвлэлийн эрх зөрчих болон түүнтэй холбоотой хууль эрх зүйн маргааны тохиолдлуудыг програм хангамж ашиглан тодорхойлоход оршино. Энэ нь хэний тал зөв эсвэл хэний тал буруу вэ гэдгийг хангалттай нотлох баримтгүйн улмаас хэн буруутай гэдгийг шийдэхэд гарах хүндрэлтэй асуудлыг шийдвэрлэхэд зориулагддаг. Иймээс, энэ тохиолдолд хэргийг хэрхэн шийдвэрлэх тухай асуудал урган гарч байна. Зохиогч танилт нь зохиол/бүтээлийн хулгайн асуудал, зохиогчийн маргааныг шийдвэрлэхэд маш чухал үүрэг гүйцэтгэдэг. Энэхүү судалгааны ажилд Монгол текст дээр шийдвэрийн модны аргыг ашиглан текстийг агуулгаар нь ангилах туршилтыг хийсэн. Бид 13 ангилал тус бүрээр сургалтын корпус үүсгэсэн. Өнөөг хүртэл Монгол текст дээр энэхүү аргыг туршсан судалгааны ажил хийгдэж байгаагүй тул шинэлэг сэдэв юм.**

Түлхүүр үгс – Машин сургалт, Текстийн ангилал, Шийдвэрийн модны ангилагч, ID3 алгоритм

I. УДИРТГАЛ

Зохиогч танилт нь зохиогчийн холбогдолтой мэдээлэл эрж хайх болон цахим хэл шинжлэл гэх мэт олон салбарт чухал ач холбогдолтой асуудал юм. Гэвч, хууль зүй болон сэтгүүл зүйн салбарт тухайн бичиг баримтын зохиогч танилт нь (жишээ нь санамсаргүй тэмдэглэл гэх мэт) тэдний амь насыг аврах боломжийг бүрдүүлж ч болох юм. Энэхүү асуудлыг шийдэх хамгийн түгээмэл тогтолцоо нь: сонгож авсан зохиогчдын жижиг, тодорхой хязгаартай багц баримт бичигт үндэслэн, тухайн бичвэрийг хэн бичсэн тухай бид тодорхойлох явдал юм. Гэхдээ энэ нь маш хүнд ажил байж болно. Иймэрхүү асуултанд хариулж чадах мэргэжлийн шүүхийн шинжээчийн бодит туршлага хэрэгтэй болно.

Өнөө үед гол нийтлэг текст ангиллын аргууд нь Байесийн ангилалын алгоритм, Шийдвэрийн мод, Хамгийн их Энтропийн загварчлал ба К-хамгийн ойр хөршийн ангилал зэрэг болно.

Энэхүү судалгааны ажилд, бид Монголын уран зохиолыг зохиогчдоор нь Шийдвэр модны аргыг ашиглан ангилна.

Бид зөвхөн Монгол улсад нэртэй романууд, тэдгээрийн зохиогчдыг сонгосон болно.

Монголд алдартай дараах 13 зохиолчдыг ангилал болгон сонгож авлаа. Үүнд:

1. Ренчин.Б
2. Лодойдамба.Ч
3. Намдаг.Д
4. Эрдэнэ.С
5. Гаадамба.Ш
6. Гармаа.Д
7. Батбаяр.Д
8. Догмид.Б
9. Жаргалсайхан.С
10. Банзрагч.Н
11. Энхболд.Д
12. Анударь.С
13. Шажинбат.Ш

Зохиолчдын ангилалын корпусаа дээрхи зохиолчдын зохиол дээр тулгуурлан үүсгэлээ.

II. ШИЙДВЭРИЙН МОДНЫ АНГИЛАГЧ АЛГОРИТМУУД

Шийдвэрийн мод нь хамгийн өргөн хэрэглэгддэг индуктив сургалтын аргын нэг юм. Энэхүү арга нь 60-аад оноос өнөөг хүртэл боловсруулагдаж байна. Аттрибутын утгыг ашигласнаар тестийн хамаарлыг тодорхойлж болно. Шийдвэрийн мод нь шийдвэр гаргах явдлыг дэмжих шинж чанарууд бүхий цэгүүдийн циклгүй граф юм. Модны салаа мөчрүүд нь зангилаануудын хоорондох жишиг харилцааг харуулна. Салаа мөчрийн хувийн жин нь салаа мөчрийн эцэг цэгийн аттрибутын утга дахь олонлогийн элемент юм. Аттрибут нь зангилаанд хамгийн багадаа хоёр хүүтэй байна, учир нь аттрибут нь бодит аттрибутын тогтоосон утгын олонлогийн элементийн тооны хэмжээ зэрэг олон салбартай

болсон байна. Модны үндэс нь нийтлэг өвөг аттрибут бөгөөд тухайн ангилал хаанаас эхлэхийг заана. Модны блокны сүүлийн байгууламж нь классын зангилаа юм. Бүх хамааралд класс бол зөв хүү, тийм болохоор тохиолдол бүрд модны навч харгалзана.

Энэхүү модыг дараах аргуудад хамаарлаар нь ангилахад ашиглаж болдог. Үүнд: Хамгийн эхлээд, аттрибут бүрийн хувьд бодит үнэ цэнэтэй шийдвэр гарсан байх ёстой. Дараагийн зангилаа нь тухайн модны салаан дээр энэхүү утгаа хадгалж байх ёстой. Хэрэв энэ зангилаа нь аттрибут бол энэ процедурыг дахин хийх ёстой. Хэрэв энэ нь класс бол мэдээлэл хадгалах шийдвэрийн модны түвшинд хүрсэн байна. Энэхүү классын шийдвэр нь эдгээр аттрибутын утга бүхий дээж нь тухайн үед хийгдсэн байх ёстой шийдвэр юм.

Үүнээс үзэхэд, өндөр утгатай өгөгдөл болон аль нэгийг сонгох хэллэгийг сурах чадавхи нь бичиг баримт ангилахад тохиромжтой. Хамгийн их танигдсан шийдвэр модны алгоритмийн нэг нь ID3 болон түүний залгамжлагч C4.5 болон C5.1 юм. Эдгээр нь рекурсивээр ангилагч шийдвэрийн модыг бүтээдэг дээрээс доош хийгддэг арга юм. Модны бүх түвшинд, ID3 хамгийн их мэдээллийн үр ашиг тодорхойлдог аттрибутыг сонгодог. Бид судалгааны ажилдаа, олон боломжуудтай ID3 шийдвэрийн модны багцыг сонгосон.

A. ID3 алгоритм

Ангилалын загварыг хөгжүүлэхэд зориулсан ID3 алгоритмийг Quinlan танилцуулж, түүнээс хойш өгөгдлийн шийдвэрийн мод гэж нэрлэгддэг болсон.

Бидэнд хэд хэдэн бичлэг бүхий олонлог өгсөн гэж үзье. Бичлэг бүр ижил бүтэцтэй бөгөөд аттрибут утгын хос бүхий тоонуудаас бүрдэнэ. Эдгээр аттрибутуудын нэг нь бичлэгийн ангилал юм. Шийдвэрийн модны шийдвэрлэх асуудал нь тухайн асуултуудад өгсөн хариулт дээр үндэслэж ямар нэгэн ангилалд ангилагдаагүй аттрибутуудыг зөв таамаглаж ямар нэгэн ангилалд хамааруулах шийдвэрийн модыг тодорхойлох явдал юм. Ихэвчлэн ангилалын аттрибут нь зөвхөн{үнэн, худал}, эсвэл {амжилт, алдаа}, эсвэл ямар нэг зүйл ижил гэсэн утга авдаг. Ямар ч тохиолдолд, түүний утгын аль нэг нь алдаа байна гэсэн үг юм.

ID3 алгоритмийн цаана байгаа үндсэн санаанууд нь:

- Шийдвэрийн модны зангилаа бүр нь аттрибутын боломжит утгаар нь бус категорийн утга бүр нь нэг нумд харгалзана. Модны навч нь түүний үндэсний навчны замд тодорхойлсон бичлэгүүдийн категори шинж хүлээгдэж буй утгыг заана. (Энэ нь Шийдвэрийн мод гэж юу болохыг тодорхойлдог.)
- Шийдвэрийн модонд зангилаа бүрийн хувьд үндэс нь зам дээр гарсан гэж үзэхгүй аттрибутуудын дунд хамгийн их мэдээлэл нь категорийн бус аттрибуттай холбоотой байх ёстой. (Энэ нь "Сайн" гэсэн шийдвэрийн мод гэж юу болохыг тогтоодог.)
- Энтропи нь зангилааны мэдээллийг хэрхэн хэмжихэд ашиглагдана.

ID3 алгоритм нь ангилалгүй $C_1, C_2, \ldots, C_n$, аттрибутуудын олонлогийг шийдвэрийн мод зурахдаа ашигладаг. Харин ангилагддаг аттрибут нь C, бичлэгүүдийн сургалтын олонлог нь T байна.

```
function ID3(
   R: a set of non-categorical attributes,
   C: the categorical attribute,
   S: a training set);
begin
   If S is empty, return a single node with
   value Failure;
   If S consists of records all with the
   same value for
      the categorical attribute,
      return a single node with that value;
   If R is empty, then return a single node
   with as value the most frequent of the
   values of the categorical attribute that
   are found in records of S;
   Let D be the attribute with largest
   Gain(D,S)
   among attributes in R;
   Let {dj| j=1,2, .., m} be the values of
   attribute D;
   Let {Sj| j=1,2, .., m} be the subsets of
   S consisting respectively of records with
   value dj for attribute D;
   Return a tree with root labeled D and
   arcs labeled d1, d2, .., dm going
   respectively to the trees
   ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..,
   ID3(R-{D}, C, Sm);
end ID3;
```

B. C4.5 алгоритм

C4.5 нь ID3-н адил аргаар сургалт мэдээллийн олонлогт үндэслэн мэдээллийн энтропи бүхий шийдвэрийн модыг бүтээдэг. Сургалтын өгөгдөл нь $S = S_1, S_2, \ldots$ олонлогоос бүрдэнэ. S_i бүр нь хэмжээст вектор $x_{1,i}, x_{2,i}, \ldots, x_{p,i}$ –ээс бүрдэнэ. x_j нь аттрибут эсвэл классын S_i алдаанд суурилна.

Модны зангилаа бүрийн үед C4.5 нь хамгийн үр дүнтэй ба нэг ангилал болон бусад баяжуулсан дэд сүлжээний дээж нь олонлогт хуваасан байгаа өгөгдлийн аттрибутуудаас сонгоно. Хуваах шалгуур нь хэвийн мэдээллийн ашиг (Энтропийн ялгаа) юм. C4.5 алгоритм нь жижиг дэд жагсаалтанд рекурсивээр ханддаг.

Энэ алгоритмд хэд хэдэн үндсэн тохиолдлууд байдаг. Үүнд:

- Жагсаалтад байгаа бүх дээжийг ижил ангилалд хамааруулна. Энэ тохиолдолд, энэ нь ердөө л тухайн ангилалыг сонгох гэж шийдвэрийн модны навчны цэгийг бий болгож байна.
- Боломжуудын аль нь ч ямар нэгэн мэдээлэл олж болдог. Энэ тохиолдолд, C4.5 нь дээд ангилалын хүлээгдэж буй утгыг ашиглаж энэ модны шийдвэрийн цэгийг бий болгож байна.
- Өмнө нь дайралцаагүй ангилалын шат тулгарсан үед C4.5 нь өндөр утгыг ашиглаж энэ модны шийдвэрийн цэгийг бий болгож байна.

III. ID3 АЛГОРИТМ АШИГЛАСАН ТУРШИЛТЫН ҮР ДҮН

Зохиолчийн 13 ангилал тус бүрээр зохиолуудыг цуглуулж, нийт 350 зохиол бүхий өгөгдөлтэй сургалтын корпус үүсгэлээ.

Хэрэв тэнцүү гарах боломжит магадлал нь n байгаа бол p-н магадлал нь $1/n$ байх ба мессежний дамжуулалж буй мэдээ нь:

$$-\log(p) = \log(n) \quad (1)$$

100 баримт бичиг байгаа бол $\log(350) = 8{,}45$, ба ингэснээр бидэнд мессеж бүрийг тодорхойлох 8,45 бит шаардлагатай гэсэн үг.

Ерөнхийдөө, хэрэв бидэнд өгсөн байгаа магадлалын тархалт[2] нь $P = (p_1, p_2, \dots, p_n)$, дараа нь бас P-ийн Энтропи гэж нэрлэдэг энэ хувиарлалтын дамжуулж буй мэдээлэл нь:

$$I(P) = (p_1 * \log(p_1) + p_2 * \log(p_2) + \cdots + p_n * \log(p_n)) \quad (2)$$

болно.

Жишээ нь, хэрэв P нь (0.5, 0.5) бол $I(P)$ нь 1, хэрэв P нь (0.67, 0.33) бол $I(P)$ нь 0.92, хэрэв P нь (1, 0) бол $I(P)$ нь 0 байна.

Хэрэв бичлэгүүдийн олонлог T категори нь аттрибутын утгад суурилан үл огтлолцох бүрэн гүйцэд ангиллын $C_1, C_2, \dots, C_k$ -д хуваагдах бол T $isInfo(T) = I(P)$ элементийн ангиллыг тодорхойлох мэдээлэл шаардлагатай, P нь $(C_1, C_2, \dots, C_k)$-н хуваалтын магадлалын тархалт[2] нь байх үед:

$$P = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_k|}{|T|}\right) \quad (3)$$

Бидний жишээнд, бидэнд Info(T) = I(9/14, 5/14) = 0.94 байна.

Хэрэв бид эхний T хуваалт дээр ямар нэг ангилалд хамаарагдахгүй аттрибут X-г T1, T2, .., Tn олонлог руу оруулвал, дараа нь T элементийн ангиллыг тодорхойлох шаардлагатай мэдээлэл Ti гэх мэт элементийн ангилал тодорхойлох хэрэгтэй мэдээллийн жигнэгдсэн нь өөрөөр хэлбэл Info(Ti)-ийн жигнэгдсэн дундаж нь:

$$\text{Info}(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{T} * Info(T_i) \quad (4)$$

Бидний жишээн дээр, зохиол аттрибутын хувьд

Info(Novel,T) = 5/14*I(2/5,3/5) + 4/14*I(4/4,0) + 5/14*I(3/5,2/5)= 0.694

Gain(X,T)-н тодорхойлсон хэмжээг авч үзвэл:

$$\text{Gain}(X,T) = \text{Info}(T) - \text{Info}(X,T) \quad (5)$$

Дээрхи нь T элементийг тодорхойлоход шаардлагатай мэдээлэл болон аттрибут X-н утгын дараагаар T элементийг тодорхойлоход шаардлагатай мэдээллийг олж авах хоёрын хоорондох гол ялгааг харуулж байна. Энэ нь аттрибут X-н мэдээллийн улмаас хийгдэж буй өсөлт юм.

Бидний жишээн дээр, зохиол аттрибутын хувьд өсөлт нь:

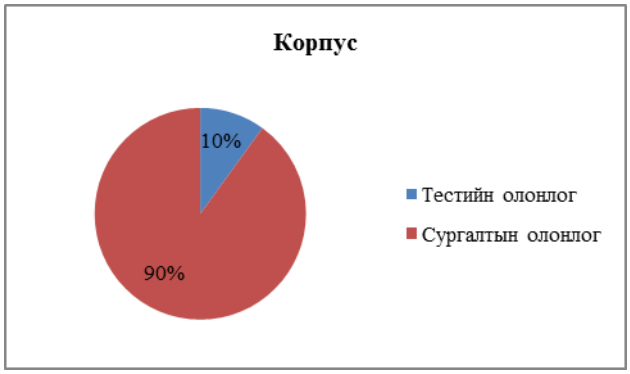
Gain(Зохиол,T) = Info(T) - Info(Зохиол,T) = 0.94 - 0.694 = 0.246.

Хэрэв бид үүний оронд Намдаг аттрибутыг авч үзвэл, Info(Намдаг,T) нь 0.892 байх ба Gain(Намдаг,T) нь 0.048 болно.

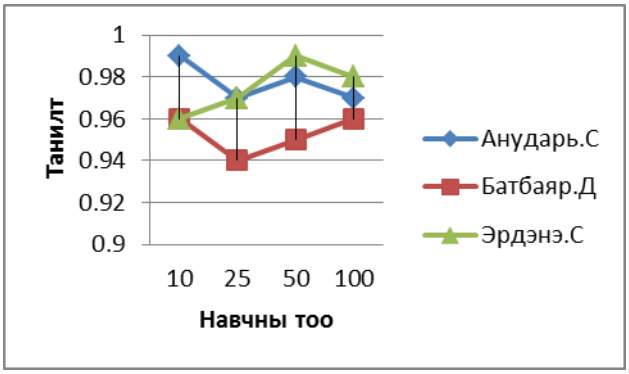
Бид аттрибутуудыг эрэмбэлэх ба цэг бүр дээр нь бас эх зам дээр гарсан гэж үзэхгүй аттрибутууд дотроос хамгийн их ашиг олох аттрибут нь байрлаж байгаа шийдвэрийн модыг үүсгэж, түүнийг ашиглаж болно.

Жижиг шийдвэрийн мод байгуулсанаар бичлэгүүд нь зөвхөн цөөн хэдэн асуултын дараа тодорхойлогдоно.

Зураг 1-д тестийн олонлог болон сургалтын олонлогийг харуулж байна.



Зураг.1. Корпусын загвар



Зураг 2. ID 3 алгоритм ашигласан танилт

TABLE I. ХҮСНЭГТ I. АНГИЛАЛЫН ТАНИЛТ

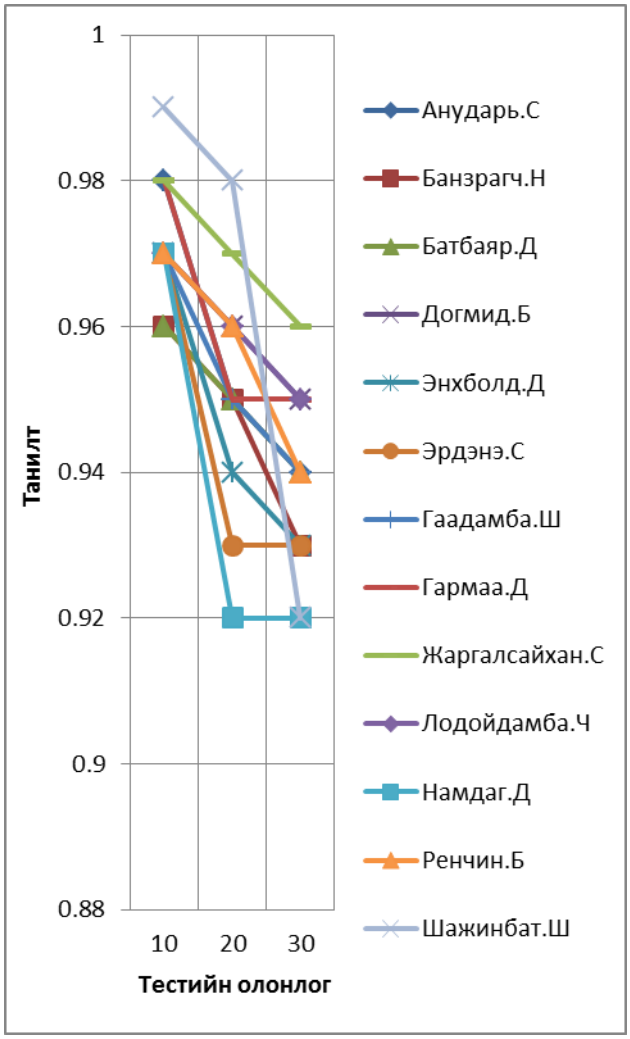
№	Зохиогч	ID 3 алгоритмийн танилтын хувь
1.	Анударь.С	0,96
2.	Банзрагч.Н	0,97
3.	Батбаяр.Д	0,96
4.	Догмид.Б	0,98
5.	Энхболд.Д	0,97
6.	Эрдэнэ.С	0,96
7.	Гаадамба.Ш	0,97
8.	Гармаа.Д	0,97
9.	Жаргалсайхан.С	0,99
10.	Лодойдамба.Ч	0,99
11.	Намдаг.Д	0,97
12.	Ренчин.Б	0,98
13.	Шажинбат.Ш	0,97
	≈ Дундаж	**0,973**

ID3 алгоритмаар Монголын зохиолийг зохиолчдоор нь ангилахад модны навчны тооноос хамаарч танилтын хувь нь өөрчлөгдөж байсан. Энэхүү харьцуулалтыг Зураг 2-оос харна уу. Бид зөвхөн 0,94-өөс дээш танилт хийсэн зохиолчдыг сонгож авсан болно.

IV. ДҮГНЭЛТ

Энэхүү судалгааны ажилд бид ID3 алгоритмийг ашигласан бөгөөд дараах үр дүнд хүрч байна. Үүнд:

- Зураг 3-т Монголын зохиолчдоор ангилсан үр дүнг харуулж байна. Корпуст нийт 13 зохиолч бүрээр 350 өгөгдөл бүхий тестийн олонлог үүсгэсэн.



Зураг 3. ID3 алгоритм ашигласан туршилт нь Монголын зохиолчдын зохиолын сургалтын олонлогоос хамаарч буй үр дүн

- Зураг 3-т бид, зохиолчоор ангилахад тестийн олонлог болон сургалтын олонлогийн өгөгдлөөс танилтын хувь нь хамаардаг гэдгийг харуулж байна.
- Бидний үүсгэсэн корпусын өгөгдөлтэй ID3 алгоритм нь сайн таарч ажиллаж байсан бөгөөд түүний давуу тал нь бага гүнтэй шийдвэрийн модыг үүсгэснээр бидэнд энэхүү алгоритмийг ашиглан илүү том, илүү нарийн өгөгдлийн олонлогтой ажиллах боломжийг олгосон. ID3 алгоритм нь анх боловруулагдсанаас хойш одоо түүний сайжруулсан хувилбар болох C4.5 гэж гарсан болно.
- Цаашид энэхүү судалгааны ажил нь, бидний санал болгож буй аргаар өргөжих, хэрэгжүүлж болохуйц байдлыг харуулахын тулд илүү их зохиогчид, мессэж зэргийг өгөгдөлдөө нэмж оруулах замаар улам өргөжих болно. Мөн илүү олон хууль бус мессежүүдийг бид судалгааны туршилтандаа тусгана. Одоогийн хандлага нь, тухайлбал яруу найраг, жүжиг гэх мэт бусад уран зохиолын холбогдолтой материалыг зохиогч бүр дээр нь дүн шинжилгээ хийх байдлаар өргөтгөх боломжтой юм.
- Өөр нэг илүү их тулгамдсан ирээдүйн чиглэл нь, автоматаар тухайн өгөгдөлд зориулсан тохиромжтой бөгөөд хамгийн оновчтой онцлог олонлогийг бий болгох явдал. Бид судалгааны ажлынхаа илүү сайн гүйцэтгэлийг цаашид төрөл бүрийн мэдээллийн санг эрлийзжүүлэн гаргах болно гэдэгт бид итгэж байна.

НОМ ЗҮЙ

[1] Christopher D. Manning, Hinrich Schutze 1999. Foundations of Statistical Natural Language Processing. Second Edition. Chapter 16 Text Categorization. pp 575-610

[2] Quinlan,J.R.: C4.5: Programs for Machine Learning, Morgan Kauffman, 1993

[3] Y.H.Li, A.K.Jain, “Classification of Text Documents”, The Computer Journal, Vol. 41, No. 8, 1998

[4] Gabor Kecskemeti , “Comparison of Classification Methods by Using the Reuters Database”

[5] Lingling Yuan, “An Improved Naive Bayes Text Classification

[6] Algorithm In Chinese Information Processing”, Jiaozuo, P. R. China, 14-15,August 2010, pp. 267-269, ISBN 978-952-5726-10-7

[7] Jian mei, Wu zhang and Suge wang 2012. Grid Enabled problem solving environments for Text Categorization

[8] Mary Slocum, "Decision making using id3 algorithm", Rivier academic journal, Volume 8, Number 2, Fall 2012

[9] Processing and Machine Translation, Chapter 1.2.1 Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development

[10] Zolboo Damiran, Khuder Altangerel, “Text Classification Experiments on Mongolian Language”, IFOST, 2013