

Зохиолын хулгайг илрүүлэх асуудалд

Б.Мөнхбуян

Мэдээллийн технологийн салбар,
Мэдээлэл, холбооны технологийн сургууль, ШУТИС
Улаанбаатар хот, Монгол улс
munkhbuyan@must.edu.mn

Хураангуй — Зохиолын хулгайг илрүүлэх орчин цагийн шаардлага, монгол улсын хууль эрх зүйн орчинг дүгнэлээ. Зохиолын хулгайн нэг төрөл болох бичвэр зохиолын хулгайн хэлбэрүүд болон тэдгээрийг таних аргуудын тухай өгүүллээ. Зохиолын хулгайг илрүүлдэг зарим онлайн системийн боломжуудыг үзүүлэв. Зохиолын хулгайг илрүүлэх монгол системийг бүтээх төлөвлөгөө боловсруулах, шаардлагатай сангуудыг тодорхойлохыг зорилго. Мөн ашиглах арга техникүүдийг санал болгов.

I. ОРШИЛ

Интернет, мэдээллийн технологи эрчимтэй хөгжиж бичиг баримтын онлайн эх үүсвэрүүд бий болсоноор их дээд сургууль, эрдэм шинжилгээний байгууллагуудад зохиолын хулгай гарах болсон [1]. Өндөр хөгжилтэй гадаадын олон улсад интернетийн хөгжил эхэлсэн үеэс эхлэн мэдээллийн технологи ашиглан зохиолын хулгайг зогсоох оролдлого хийж эхэлсэн [2]. Харин монгол улсад мэдээллийн технологи ашиглан зохиолын хулгайг илрүүлэх судалгаа болон програм хангамж хараахан хийгдээгүй байна. Зөвхөн эрдэмтэн судлаачид, багш нар болон програм зохиогчид Монголын улсын зохиогчийн эрх болон түүнд хамаарах эрхийн тухай хуулийн хүрээнд л хянаж, шалгаж байна. Ингэж мэдээллийн технологи ашиглахгүйгээр шалгах нь түвэгтэй бөгөөд цаг хугацаа, хөрөнгө мөнгө их шаардсан ажил байдаг. Иймээс мэдээллийн технологи ашиглан зохиолын хулгайг таних шаардлага гарч байна.

Монгол улсын зохиогчийн эрх болон түүнд хамаарах эрхийн тухай хуульд дараах чухал зүйлсийг оруулсан байна [3]. Үүнд:

1. Хуулийн 7 дугаар заалтад шинжлэх ухаан, утга зохиолын бичмэл бүтээл болон компьютерийн програмыг зохиогчийн эрхээр хамгаална гэсэн байдаг. Иймээс дараа дурьдах хоёр төрлийн зохиолын хулгайг мөн монгол улсад зохиолын хулгайд тооцдог байна.
2. Хуулийн 12 дугаар заалтад зохиогч өөрийн бүтээлийг бүх хэлбэрээр ашиглах эрхүүдийг эдлэхээс гадна бусдад хуулбарлах, нийтлэх, нийтэд түгээх, үүсмэл бүтээл болгон өөрчлөх, түрээсээр ашиглуулах эсэхийг зөвшөөрлийн дагуу олгож болно.
3. Хуулийн 24 дүгээр заалтад эшлэл авах болон эрдэм шинжилгээ, судалгааны бүтээл, шүүмжид эшлэл авч хэсэгчлэн ашиглахыг зохиогчийн эрхийн зөрчилд тооцохгүй гэсэн байдаг.

Британика нэвтэрхий тольд зохиолын хулгайг дараах байдлаар томъёолсон байдаг [4]. Зохиолын хулгай бол өөр хүний бичвэрээс өөрийн мэт авч бүтээл бичих үйл явц юм. Энэ нь зохиогчийн эрхийн хуулийн зөрчилд хамаарна гэжээ.

Монгол улсын зохиогчийн эрх болон түүнд хамаарах эрхийн тухай хууль болон олон улсын Бернийн конвенциос үзвэл хоёр төрлийн зохиолын хулгай байна [5]. Эхний төрөл нь програмын кодын хулгай юм [6]. Програмын кодын хулгай буюу програмчлалын бүтээлийн хулгайг их дээд сургуульд оюутнууд, програм хангамжийн компанид програм зохиогчид үйлддэг байна. Энэ нь анхны зохиогч нарт мэдэгдэхгүйгээр, хууль бусаар, түүний бичсэн програмын эх кодыг бүтнээр нь болон хэсгээр хуулах эсвэл бага хэмжээгээр өөрчлөн хуулах болон өөр програмчлалын хэл рүү хөрвүүлэн ашиглахыг хэлнэ. Дараагийн зохиолын хулгайн төрөл бол бичвэр зохиолын хулгай юм. Энэ төрлийн зохиолын хулгайг их дээд сургууль, эрдэм шинжилгээний байгууллагад оюутнууд, судлаачид үйлддэг. Цаг хугацааны хувьд бичвэр зохиолын хулгайг таних хэрэгцээ нэн түрүүнд шаардлагатай тул энэхүү судалгааны ажилд бичвэр зохиолын хулгайг таних тухай өгүүллээ.

II. Зохиолын хулгайн хэлбэр

Судалгаанаас үзвэл бичвэр зохиолын хулгайн дараах хэлбэрүүд байна. Үүнд:

1. шууд хуулсан (copy past) - өөрийн бүтээл мэт өөрчлөлтгүй хуулбарлан буулгах [7];
2. эшлэл будлиулсан (aggregation) – анхдагч өгүүллийг биш, өөр өгүүллийг эшлэл авах [8];
3. үг сольсон (word switch) – үгийг ижил утгатай эсвэл өөр үгээр солих [7];
4. гишүүн өгүүлбэрийн байрлал сольсон (sentence rearrangement) – нийлмэл өгүүлбэр дэх өгүүлбэрүүдийн байрлалыг солих [1];
5. өгүүлбэр зүйрлэсэн (metaphor) - ижил утгатай өөр өгүүлбэр болгох [7];
6. холимог (fusion) - олон эх үүсвэрээс хуулбарлаж, шинэ холимог үүсгэх [8];
7. зохион байгуулалтыг хуулсан (organizational) - өгүүллийн ерөнхий зохион байгуулалтыг хуулбарлах [1];
8. хэлбэр өөрчилсөн (style) – өгүүллийн үгсийг сольж, тайлбар болон санаануудын дарааллыг шууд хуулах [7];

9. санаа хуулсан (idea) - бичвэрийг хуулбарлалгүйгээр үндсэн зарчмыг авах [7];
10. орчуулсан (translation) – эшлэл авахгүйгээр өөр хэл рүү орчуулах [9];

Дээрх зохиолын хулгайн хэлбэрүүдээс 1- ээс 6 дугаар хэлбэрийн буюу шууд хуулсан, эшлэл будлиулсан, үг сольсон, гишүүн өгүүлбэрийн байрлал сольсон, өгүүлбэр зүйрлэсэн, холимог зохиолын хулгайн хэлбэрүүд нь өгүүлбэр, догол мөр, бүлгийн хэмжээнд үйлдэгддэг тул хэсэгчилсэн зохиолын хулгайн хэлбэр гэж нэрлэв. Харин 7- оос 10 дүгээр хэлбэрийн буюу зохион байгуулалтыг хуулсан, хэлбэр өөрчилсөн, санаа хуулсан, орчуулсан зохиолын хулгайн хэлбэрүүд нь бүтэн өгүүлэл, зохиол бүтээлийн хэмжээнд үйлдэгддэг тул бүтэн зохиолын хулгайн хэлбэр гэж нэрлэв. Гэхдээ 10 дугаар хэлбэр буюу орчуулсан зохиолын хулгай нь илрүүлэхэд тун төвөгтэй хэлбэр юм. Мөн 2 дугаар хэлбэрийн буюу эшлэл будлиулах зохиолын хулгай нь бусад хулгайн хэлбэрээс эрс ялгаатай бөгөөд эшлэл зөв эсэхийг шалгах асуудал байдаг. Энэ хэлбэрийн зохиолын хулгай нь 1 дүгээр хэлбэрийн шууд хуулах зохиолын хулгайтай нэг ижил байх боломжтой. Иймээс есөн хэлбэрийн зохиолын хулгай байна гэж үзэж болно.

III. Зохиолын хулгайг таних аргууд

Бичвэрийн зохиолын хулгайг таних аргуудыг таван бүлэгт хувааж болно [7]. Үүнд:

1. Хэлзүйд суурилсан аргууд;
2. Утгад суурилсан аргууд;
3. Хэлзүй, утга хосолсон аргууд;
4. Эх үүсвэрийг тодорхойлох аргууд;
5. Ангилалд хуваах аргууд;

1. Хэлзүйд суурилсан аргууд

Энэ төрлийн арга нь бичиг баримтын дүрмийн бүтцэд анхаарлаа хандуулдаг. Энэ төрлийн аргууд тэмдэгт мөрд суурилсан харьцуулалт хийх замаар бичиг баримт хоорондын төсөөг хэмждэг. Зөвхөн шууд хуулах 1-р ангиллын зохиолын хулгайд тохиромжтой. Харин ижил утгатай үгээр сольсон эсвэл дахин бичих замаар өөрчилсөн бичвэрийг танихад тохиромжгүй.

2. Утгад суурилсан аргууд

Энэ төрлийн арга вектор орон зайн загвар ашиглан бичиг баримт хоорондын төсөөг тодорхойлоход анхаарлаа хандуулдаг. Хоёр бичиг баримтын бичиглэлийн онцлог шинжүүдийг тодорхойлж, хооронд нь харьцуулах замаар түүнээс төсөөг тооцоолдог. Ийм арга бичиг баримтыг бүтнээр нь ашиглаж, вектор орон зайн загвар гаргадаг. Гэвч хэсэгчлэн хуулбарласан бол энэ нь сайн үр үзүүлж чаддаггүй дутагдалтай тал байдаг.

3. Хэлзүй, утга хосолсон аргууд

Эх хэлний хувьд энэ төрлийн арга хамгийн чухал арга гэж үздэг. Энэ арга харьцангуй үр бүтээлтэй байж үр дүнг сайжруулсан. Хэлзүйд суурилсан аргад танигдахгүй, ижил утгатай үгээр солих эсвэл өөрчлөн

бичих замаар өөрчлөгдсөн бичвэрт болон хуулсан бичвэрт тохиромжтой. Мөн утгад суурилсан аргын дутагдалтай талыг шийдсэн. Энэ арга мөн утгад суурилсан аргаар тодорхойлж чадахгүй, бичиг баримтын хуулбарласан хэсгийн байрлалыг тодорхойлж, хоёр бичиг баримтын төсөөг тооцоолж чаддаг.

4. Эх үүсвэрийг тодорхойлох аргууд

Энэ төрлийн арга бичвэрийн хэсгийг хуулбарлаж болох бичиг баримтуудаас бүрдэх лавлах хөмрөгийг авч үздэг. Бичвэрийн хэсэг нь догол мөр, тогтмол хэмжээтэй багц үгс, багц өгүүлбэр гэх мэт байж болно. Сэжигтэй бичиг баримтын хэсгүүдийг лавлах хөмрөг доторх бичиг баримтууд дахь баримтын хэсэгтэй харьцуулах замаар зохиолын хулгайг шалгадаг.

Энэ төрлийн арга мэдээллийн төрөл болон агуулгын хайлттай (Information retrieval) төстэй байдаг. Энгийн алгоритм ашиглавал ажиллагааны хугацаа ихэсдэг, лавлах хөмрөгийн хэмжээ ихэсдэг тул үр дүн гарах боломжгүй байдаг. Иймээс хейшлэлт, бичиглэлийн онцлог шинжийг харьцуулах арга ашиглах нь өндөр үр дүнтэй байдаг. Зарим системд бичиг баримт эсвэл бичгийн хэсгийн агуулгыг илэрхийлсэн хэд хэдэн бичиглэлийн онцлог шинжүүд үүсгэж ашиглаж байсан. Сэжигтэй бичиг баримтын хэсгүүдийн онцлог шинжийг хейш эсвэл бичиглэлийн онцлог шинжээс бүрдэх лавлах хөмрөгөөс харьцуулж хайдаг. Өөрөөр хэлбэл онцлог шинжээр нь төсөөг хэмждэг. Энэ аргыг анх 1995 онд танилцуулж байсан. Энэ төрлийн аргыг мөн вектор орон зай Rd дахь хамгийн ойр хөршийн асуудал байдлаар тодорхойлдог.

5. Ангилалд хуваах аргууд

Бичиг баримтыг ангилах нь олон зорилгоор мэдээллийн төрөл болон агуулгын хайлтад ашиглагддаг чухал арга техникүүдийн нэг юм. Хайлтын хугацааг багасгах замаар өгөгдлийн хайлтыг сайжруулж, бичиг баримтуудыг нэгтгэн дүгнэхэд ашигладаг. Энэ нь мөн үр дүнг илэрхийлэхэд ашиглагдаж болно.

Бичиг баримтын ангилалт нь зохиолын хулгайг танихад хайлтын хугацааг багасгахын тулд ашигладаг. Гэвч одоо хүртэл ангилалд хуваах аргад зарим дутагдалтай тал болон хугацаа, хэмжээний хүндрэлүүд байсаар байна [10].

IV. Зохиолын хулгайн хэлбэр ба таних аргуудын ХОЛБОО ХАМААРАЛ

Эдгээр бичвэрийн зохиолын хулгайг таних бүлэг аргууд зохиолын хулгайн хэлбэрүүдийг таньж чадахыг дараах Хүснэгт 1-ээр үзүүлэв.

№	хулгайн хэлбэр	хэлзүйд суурилсан	утгад суурилсан	хэлзүй, утга хосолсон	эх үүсвэрийг тодорхойлох
1	шууд хуулсан				
2	эшлэл будлиулсан				
3	үг сольсон				
4	гишүүн өгүүлбэрийн байрлал сольсон				
5	өгүүлбэр зүйрлэсэн				
6	холимог				
7	зохион байгуулалтыг хуулсан				
8	хэлбэр өөрчилсөн				
9	санаа хуулсан				
10	орчуулсан				

Эдгээрээс харахад орчуулсан зохиолын хулгайн хэлбэрийг машин орчуулгын системтэй хамтарч гүйцэтгэх шаардлагатай байна. Харин ангилалд хуваах аргууд нь зөвхөн хайлтын хугацааг багасгах, хадгалах мэдээллийг багасгахын тулд ашигладаг.

Дээрх аргуудыг ашигласан PlagAware, PlagScan, iThenticate, CheckForPlagiarism.net, PlagiarismDetection.org зэрэг онлайн системүүдийг судалж үзэв. Эдгээр нь дараах боломжуудыг тусгасан байдаг [7]. Үүнд:

- өөрийн их хэмжээний өгөгдлийн сантай,
- интернетээр шалгах,
- багш, оюутнууд эрдэм шинжилгээний өгүүлэл шалгах,
- олон бичиг баримтыг шалгах,
- олон хэлний боломжтой байх,
- өгүүлбэрийн бүтэц, ижил утгаар шалгах.

Дүгнэлт

1. Интернет, мэдээллийн технологи эрчимтэй хөгжсөн өнөө үед их дээд сургууль, эрдэм шинжилгээний байгууллагуудын хүндрэл болсон зохиолын хулгайг зогсоох шаардлагатай.
2. Зохиогчийн эрх болон түүнд хамаарах эрхийн тухай хуулийг хэрэгжүүлэхэд мэдээллийн технологийг ашиглах шаардлагатай.
3. Зохиолын хулгайг илрүүлэх монгол системийг

I. Хэсэгчилсэн зохиолын хулгайн хэлбэрүүд:

- i. шууд хуулсан ба эшлэл будлиулсан,
- ii. үг сольсон,
- iii. гишүүн өгүүлбэрийн байрлал сольсон,
- iv. өгүүлбэр зүйрлэсэн,
- v. холимог,

II. Бүтэн зохиолын хулгайн хэлбэрүүд:

- i. зохион байгуулалтыг хуулсан,
- ii. хэлбэр өөрчилсөн,
- iii. санаа хуулсан,

III. Орчуулсан хулгайн хэлбэр.

4. Дээрх зохиолын хулгайн хэлбэр тус бүрийг хэрэгжүүлэхэд тодорхой сан, алгоритмуудыг зохиох шаардлагатай. Жишээлбэл:

- үг сольсон хулгайн хэлбэрийг илрүүлэхэд ижил утгатай үгийн сан шаардлагатай.
- гишүүн өгүүлбэрийн байрлал сольсон хулгайн хэлбэрийг илрүүлэхэд монгол хэлний өгүүлбэр зүйн сан шаардлагатай.

5. Бүтэн зохиолын хулгайн хэлбэрүүдийг илрүүлэхдээ хоёр бичиг баримтын төсөөг тодорхойлох математик, машин сургалтын аргуудыг ашиглана.

6. Орчуулсан хулгайн хэлбэрийг илрүүлэхийн өмнө машин орчуулгын асуудлыг шийдвэрлэсэн байх шаардлагатай.

7. Зохиолын хулгайг илрүүлэхдээ хэлзүй, утга хосолсон аргуудаас ашиглах нь харьцангуй олон зохиолын хулгайн хэлбэрүүдийг илрүүлэх боломжтой. Ингэхдээ эх үүсвэрийг тодорхойлох арга дахь бичиг баримтын бичиглэлийн онцлог шинжүүдийг тодорхойлж, тэдгээрээр лавлах хөмрөг үүсгэж, онцлог шинжээр хайхдаа хейшлэлт ашиглах шаардлагатай. Мөн хайлтын хугацааг багасгахын тулд ангилалд хуваах аргууд ашиглаж болно.

8. Эх хэлний боловсруулалтын болон компьютерийн хэлшинжлэлийн арга техникүүдийг монгол хэлний зохиолын хулгайг танихад ашиглах шаардлагатай [9].

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

бүтээхдээ цаг хугацааны хувьд нэн түрүүнд шаардлагатай байдал болон хэрэгжүүлэхэд хялбар байдлаас нь дараах дарааллаар зохиолын хулгайн хэлбэрүүдийг хамруулна.

НОМЗҮЙ

- [1] Bright Hub Education, <http://www.brighthubeducation.com/help-with-writing/39393-types-of-plagiarism-and-examples/>, (2016-04-17).
- [2] Alan Parker, James O. Hamblen, “Computer Algorithm for Plagiarism Detection”, IEEE Transactions on Education, 1989.
- [3] Зохиогчийн эрх болон түүнд хамаарах эрхийн тухай, <http://www.legalinfo.mn/law/details/283>, 2006.
- [4] Encyclopedia Britannica, <http://www.britannica.com/topic/462640/plagiarism>, (2016-04-17).
- [5] Зохиогчийн эрхийн тухай дэлхийн оюуны өмчийн байгууллагын гэрээ, <http://www.legalinfo.mn/annex/details/6350?lawid=10199>, 2002.
- [6] Shanmugasundaram Hariharan, “Automatic Plagiarism Detection using Similarity Analysis”, The International Arab Journal of Information Technology, 2012.
- [7] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel, “Overview and Comparison of Plagiarism Detection Tools”, Dataso 2011, 2011.
- [8] Sonawane Kiran Shivaji, “Plagiarism Detection by using Karp-Rabin and String Matching Algorithm Together”, International Journal of Computer Applications, 2015.
- [9] Angry Ronald Adam, Suharjito, “Plagiarism Detection Algorithm using Natural Language Processing based Grammar Analyzing”, Journal of Theoretical and Applied Information Technology, 2014.
- [10] Jain, K. Murty, M. N., Flynn, P. J., “Data clustering: a review”, ACM Computing Surveys, 1999.